DISCOVERING NOVEL HUMAN STRUCTURAL VARIATION FROM DIVERSE POPULATIONS AND DISEASE PATIENTS

AN EXPLORATION OF WHAT HUMAN GENOMICS MISSES BY RELYING ON REFERENCE-BASED ANALYSES

by
Rachel Michelle Sherman

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
January 2021

Abstract

Since the completion of the human genome project, the field of genomics has relied on the human reference genome for nearly all analyses. Population genetics, disease association studies, and beyond all begin by comparing an individual's sequenced genome to the human reference. However, the human reference genome is not only still incomplete, but also not an accurate representation of humanity; it is derived primarily from a single individual, and cannot possibly represent the scope of human diversity. By using this genome as a template, we bias our studies. In this thesis we examine large regions of structural variation between individuals that are often missed by comparing solely to the human reference genome. We use multiple strategies to uncover variation, including performing localized assembly on whole genome sequencing reads not matching the reference genome from 910 individuals of African ancestry, and utilizing new, long-read sequencing technologies in disease patients. We demonstrate that vast amounts of sequence present in human populations, nearly 300 megabases in the case of the African ancestry dataset, are missing from the reference genome, as well as that many non-reference sequences are present in breast cancer and Mendelian disease patients, which could have yet-to-be-discovered disease relevance. We find evidence of novel non-reference sequences which are genic and transcribed in many individuals, which may have functional relevance. Finally we present strategies for integrating the wealth of short-read sequencing data currently available with the limited but growing number of newer, long-read sequenced samples to gain new insights previously inaccessible using short-read data alone.

Committee: Steven Salzberg (Primary Advisor), Michael Schatz, Ben Langmead, Winston Timp

Acknowledgments

The research in this thesis could not have happened without my advisor, Steven Salzberg, nor without Mike Schatz, who has been not just a research advisor but a mentor since my days as a summer undergrad at Cold Spring Harbor Lab. My own summer undergraduate students, Juliet Forman and Will Cho, were invaluable in helping produce work for this thesis. Sergey Aganezov and Melanie Kirsche worked jointly with me on several pieces of this work, and I had many productive research discussions with them, as well as with many members of the Salzberg, Schatz, and Langmead labs who were always happy to talk through research problems or ideas.

But research is only one part of completing a PhD, and I think I would have left Baltimore PhD-less years ago without the communities I've found here. The women of GRACE have provided invaluable emotional support over the years, especially Kate Fischl and Huda Khayrallah. I've grown to love Baltimore thanks to my volunteer family at the MD SPCA, who brought me to my wonderful, beautiful, and completely insane rescue pit bull Misty, and to Rae Borsetti, who went from my MD SPCA manager, to a friend, to the person who properly introduced me to Baltimore. To my housemate and closest friend, Elissa Sutlief, and to Matt Hachi Reid: our little pandemic family has made crossing the finish line bearable in this otherwise strange and terrible year. And I owe so much to the friends that are always there for me, even when they're far away: Meredith Sturmer, Coline Devin, and Michelle Chesley, who I know I can always call when I need them. And finally, thanks to my parents, Bruce and Helene Sherman, for their unwavering love and support. I could not have made it here without you.

in memory of my mom
Helene Sherman z”l

who taught me
to stay strong
and persevere

but that knowing
when to stop
is just as important

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 1: Introduction to human whole genome sequencing

Sections of Chapter 1 have been previously published as:

Sherman, R. M. & Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews Genetics*, 21, 243–254.

1.1 The human reference genome

Much of the field of genomics revolves around the existence of reference genomes, which are roadmaps for a 'typical' individual of each species. The creation of each reference was, and still remains, a major focus of the genomics community, with 13 years and US$2.7 billion[1] spent on the creation of the human reference genome alone. An initial draft of the human reference genome was first published in 2001[2,3]. The genome consisted of sequence from approximately 20 individuals, who answered an advertisement for volunteers in the *Buffalo News*, a newspaper in Buffalo, New York, USA. To sequence these individuals, DNA was extracted from a blood sample, and was sheared into ~150–200 kb pieces, which were inserted into bacterial artificial chromosomes (BACs) to be sequenced. This approach meant that each ~150 kb segment could be sequenced and assembled separately, reducing errors caused by ubiquitous repeats that occur throughout the genome. Furthermore, a physical map of the genome was created to determine the relative locations of the BAC clones along the chromosomes. Thus the human reference genome was assembled as a mosaic of these sequenced individuals, where one BAC-length segment might come from one individual, and the next segment from a different individual, and so on. The individuals who provided the DNA were anonymous.

The original version of the human reference genome contained 2.69 Gbp and nearly 150,000 gaps -- regions where the sequence was not able to be resolved. The genome has undergone many major updates since 2001 to produce the current version, GRCh38.p13, which contains 2.95 Gbp of sequence and only 349 gaps[4]. These updates have included filling in gaps where no sequence was present, replacing rare alleles in the genome with the more common variants, and adding alternative sequences representing divergent variants of some portion of the reference genome, although these alternative sequences are often not considered by analysis pipelines, and in some cases can confound downstream analyses. However, the underlying genetic background of the current human reference remains the same as in the initial version — a mosaic of sequences from a small number of anonymous individuals.

In 2010, a paper describing the Neanderthal genome additionally performed an analysis on the human reference (version GRCh37)[5]. That analysis used the original BAC information to trace which anonymous donor was the source for each segment of the genome, and then used population-specific single nucleotide polymorphisms (SNPs) to determine the ancestry of each donor. This process revealed that ~2/3 of the reference genome sequence was comprised of DNA from one male donor with the anonymous identifier RPCI-11, and that RPCI-11 was almost certainly 50% African and 50% European. The analysis examined both the full makeup of BAC clones and inferred ancestral makeup of GRCh37 (Figure 1.1), as described in their supplemental material[5] (Green *et al* supplement, p 146). Because scientists continue to use the human reference genome as a baseline for nearly all human genetics studies, it is important to acknowledge that it does not represent the whole population. Rather, it is a mixture of

ethnicities, predominantly sequence from a European/African admixed individual. Furthermore, as a mosaic of many individuals, it may not represent variant combinations that exist in any individual.

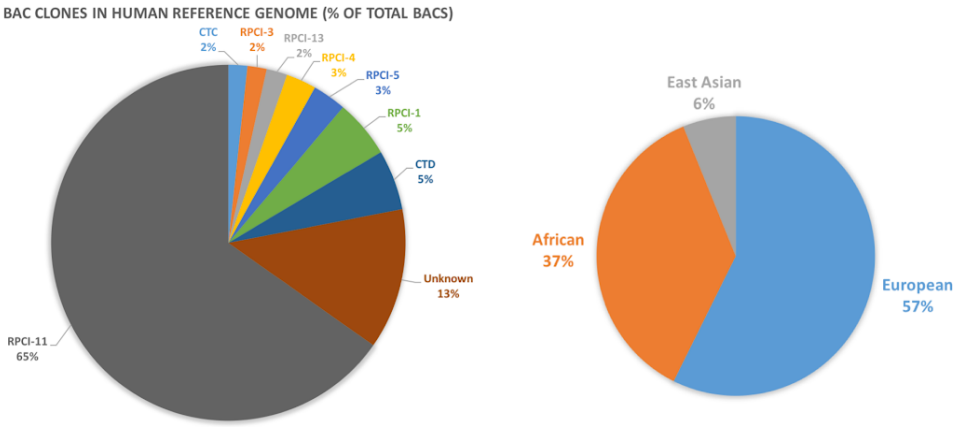


Figure 1.1 | Ethnic makeup of GRCh37. Left: The estimated breakdown of the percentage of BAC clones derived from each original donor. Right: The inferred ancestry makeup of GRCh37. All numbers were taken from Green et al 2010, supplement.

Today, most human whole-genome sequencing analyses begin by aligning sequencing reads to the human reference genome. The existence of a reference eliminates the need for *de novo* assembly, where reads must be overlapped and pieced together to create the genome sequence from scratch. Not only does *de novo* assembly require high coverage data and high computational expense, but due to short read lengths and long repeats in the human genome, *de novo* assemblies are typically broken into many pieces (contiguous sequences called 'contigs') where the contigs cannot be merged together unambiguously either due to repetitive content, or gaps in sequencing in regions which are difficult to sequence or assemble such as the centromeres. Alignment based strategies eliminate these problems, instead relying on the extensive work that has gone into producing a highly contiguous reference genome, and using

the reference as a template. Reads are lined up to the reference to determine their positions, and exact matches are not required, so SNPs and small indels can be discovered via alignment strategies (Figure 1.2).

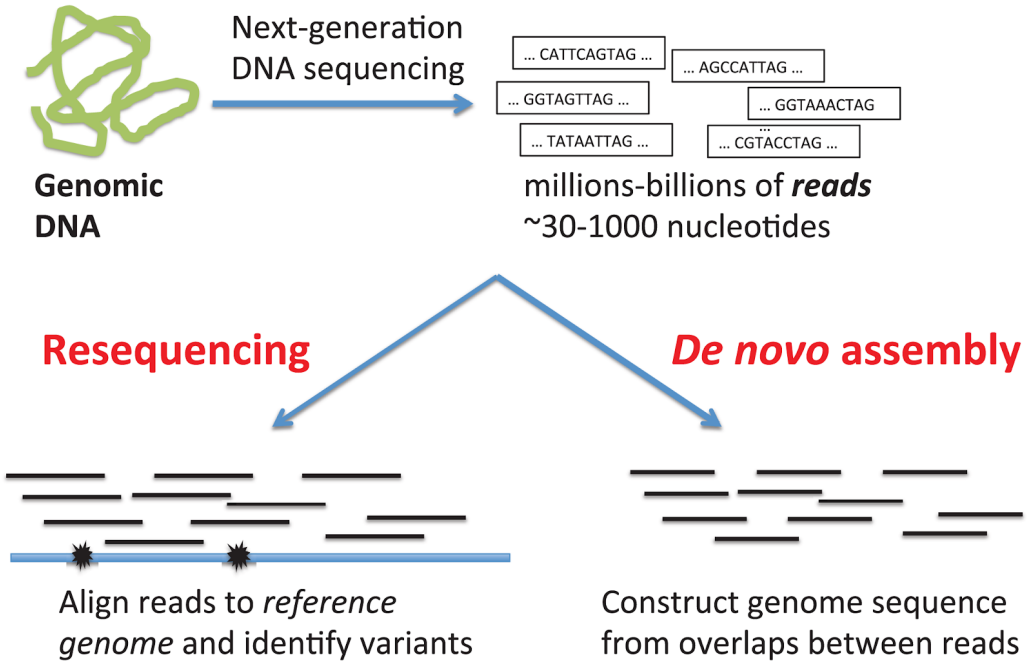


Figure 1.2 | Alignment vs de novo assembly strategies. Alignment based strategies, also called re-sequencing, align reads to a reference genome to find variants. *De novo* assembly, on the other hand, creates a genome by overlapping reads. Variants can then be found by comparing *de novo* assemblies. Figure from Raphael BJ (2012) Chapter 6: Structural Variation and Medical Genomics. PLoS Comput Biol 8(12): e1002821.

Alignment-based analysis approaches come with their own challenges, however. The alignments of the short-reads from a new individual are biased by the reference genome; alignments with the fewest mismatches will be preferred, but the "best" alignment will not necessarily always be the biologically "correct" alignment (aka where the read actually came from in the sequenced individual). Additionally, reads might align equally well to multiple locations when

repeats are present, creating ambiguity in not just alignments, but in subsequent SNP and small indel calls (Figure 1.3).

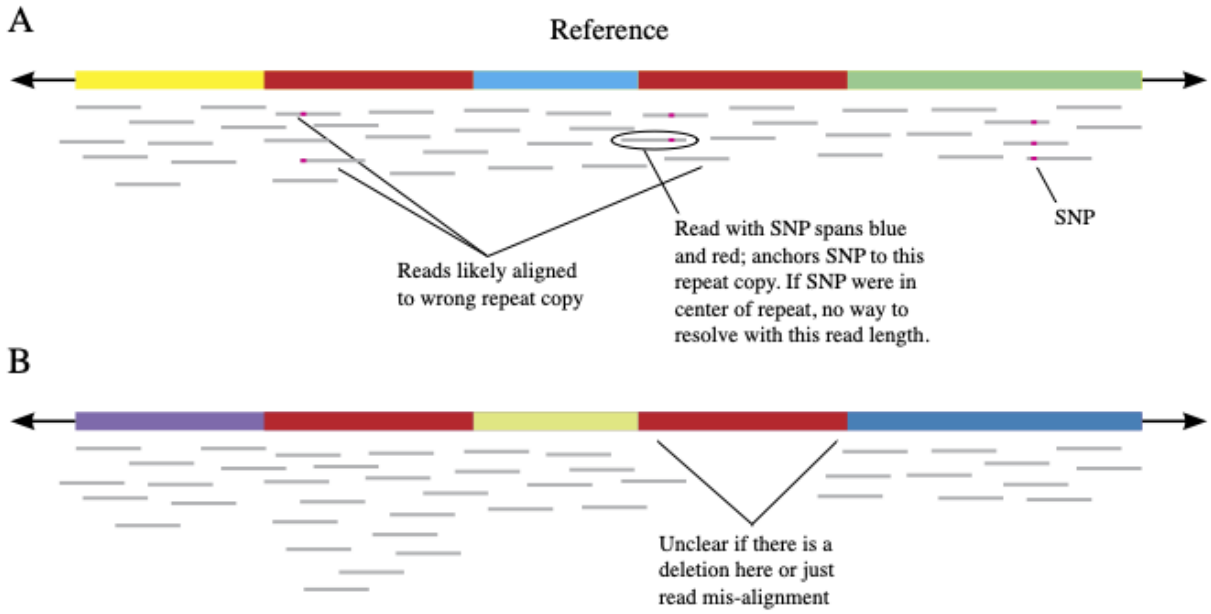


Figure 1.3 | Short read alignment ambiguity complicates even small variant calling. Examples of possible short read alignments in a region with a repeat (red). In (A), there is a straightforward SNP to call in the green region of the reference. However, there is also a SNP appearing in reads aligned to both repeat regions. This could reasonably be called as a heterozygous SNP in both regions. We might infer that this SNP likely belongs in the rightmost repeat, based on presence of a read spanning the repeat edge from blue to red, but not yellow to red. However, we still cannot rule out the possibility this is heterozygous in both repeats. A SNP centered in this repeat region would have no chance of having this ambiguity resolved via repeat edge spanning reads. In (B), the reads from the repeat happen to be aligned to only the first copy. While the coverage increase in the first copy may be detectable, this also may be a deletion of the second repeat copy; it is ambiguous.

Regions of large difference between the sequenced individual and the reference, termed structural variation, might be missed entirely, as short-reads with no match in the reference genome will be discarded in downstream analyses. Some of these problems are short-read specific, and newer third generation sequencing technologies may help overcome some of these caveats, however, even with improved alignments strategies, the reference genome

remains a biased picture of a human genome; no single genome can realistically represent the genomic diversity present across human populations.

1.2 Second and third generation sequencing technologies

Second generation, also known as next generation sequencing (NGS) or short-read (SR) sequencing technologies for whole genome sequencing, predominantly Illumina sequencing, consists of shearing the human genome into short sequences of approximately 150-250 bases in length. The length of sequencing is limited by the methodology; longer DNA fragments can be produced but a longer fragment cannot be sequenced accurately. Illumina uses a strategy called 'sequencing by synthesis' (Figure 1.4). This strategy involves adding fluorescent nucleotides onto single-stranded DNA, one at a time, and imaging; the color determines the added nucleotide and thus implies it's complement (Figure 1.4c). Each addition is considered one cycle. However, to get a strong enough fluorescence signal to image, multiple copies of the sequence must undergo this process simultaneously, so multiple nucleotides will fluoresce the same color at the same time. Thus an amplification step precedes nucleotide addition (Figure 1.4b). However, as the sequencing continues, errors such as a nucleotide not attaching in a cycle, or two attaching in a cycle, etc on one of the molecules will eventually accumulate and put the fluorescence of the copies out of sync. Once too many errors accumulate the fluorescence color (and thus nucleotide of the original strand) cannot be accurately determined, limiting the read length. To combat this, 'reads' are typically sequenced from each end of longer fragments, creating 'paired end' reads that are (approximately) some known distance apart based on the fragment size -- for example, if 150 bp are sequenced from either end of ~500bp

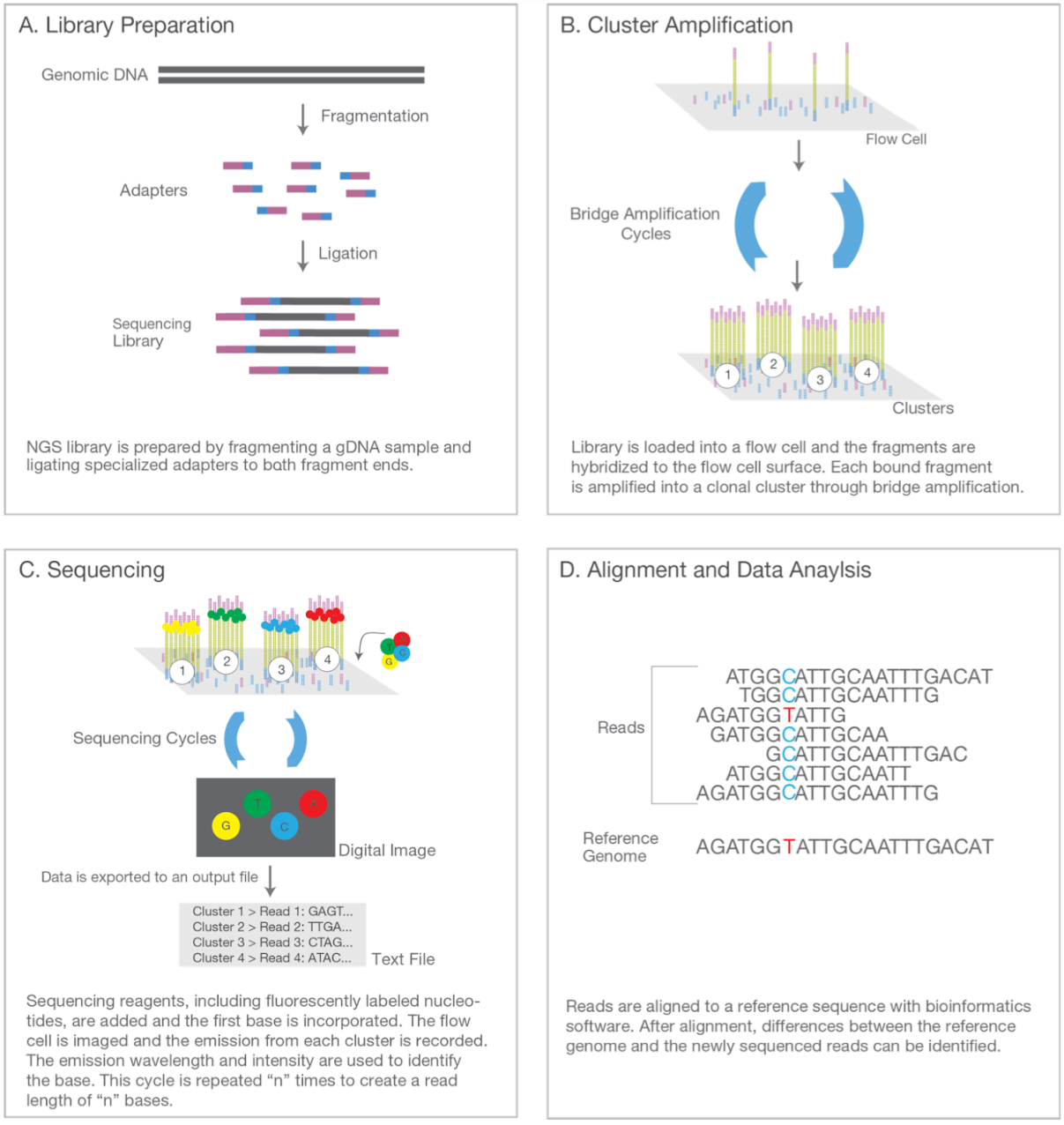


Figure 1.4 | Illumina sequencing overview. (A) Library preparation shearing and adding adapters to the DNA, (B) amplification, to create clusters of the same molecule on a 'flow cell', (C) sequencing via addition of fluorescent nucleotides and imaging, and (D) read alignment for Illumina short-read sequencing. Figure taken from Illumina's An Introduction to Next-Generation Sequencing Technology [6].

fragments, the 'reads' would be known to be ~200bp apart. This paired end strategy aids in both read mapping and assembly, by providing additional information about how reads relate to

one another spatially. Larger fragments can also be used, where the 'mate pair' reads might be ~3000 or ~5000 bases apart from one another, to provide some longer-range information [6,7].

The majority of available whole genome sequencing data is short read sequencing. Illumina sequencing is very high throughput, with the ability to sequencing many molecules and many samples simultaneously in parallel, and sequencing costs have rapidly decreased. The amount of sequencing data available in the Sequence Read Archive (SRA) is currently over 46 petabases, and still growing rapidly (https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/). However, NGS data is notoriously difficult to assemble. Short reads do not span long repeats which are highly present in human genomes (and even more so in many plants), so the order in which unique sequences between repeated sequence cannot be determined (Figure 1.5). Even with high coverage data, *de novo* assembly results in highly fragmented genomes, thus the reference genome is heavily utilized for re-sequencing experiments, and nearly all NGS analyses begin with alignment to the human reference genome, a strategy which poses other challenges (refer back to Chapter 1.1).

Recent advances in sequencing technologies have led to two distinct approaches to generating longer read lengths. The two methods are commonly referred to by the companies which developed the technologies; Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Both methods are able to produce much longer sequences, but PacBio's read lengths are limited by the methodology, whereas ONT read lengths are limited only by the DNA preparation; if DNA can be extracted and kept in large fragments stably, they can be sequenced.

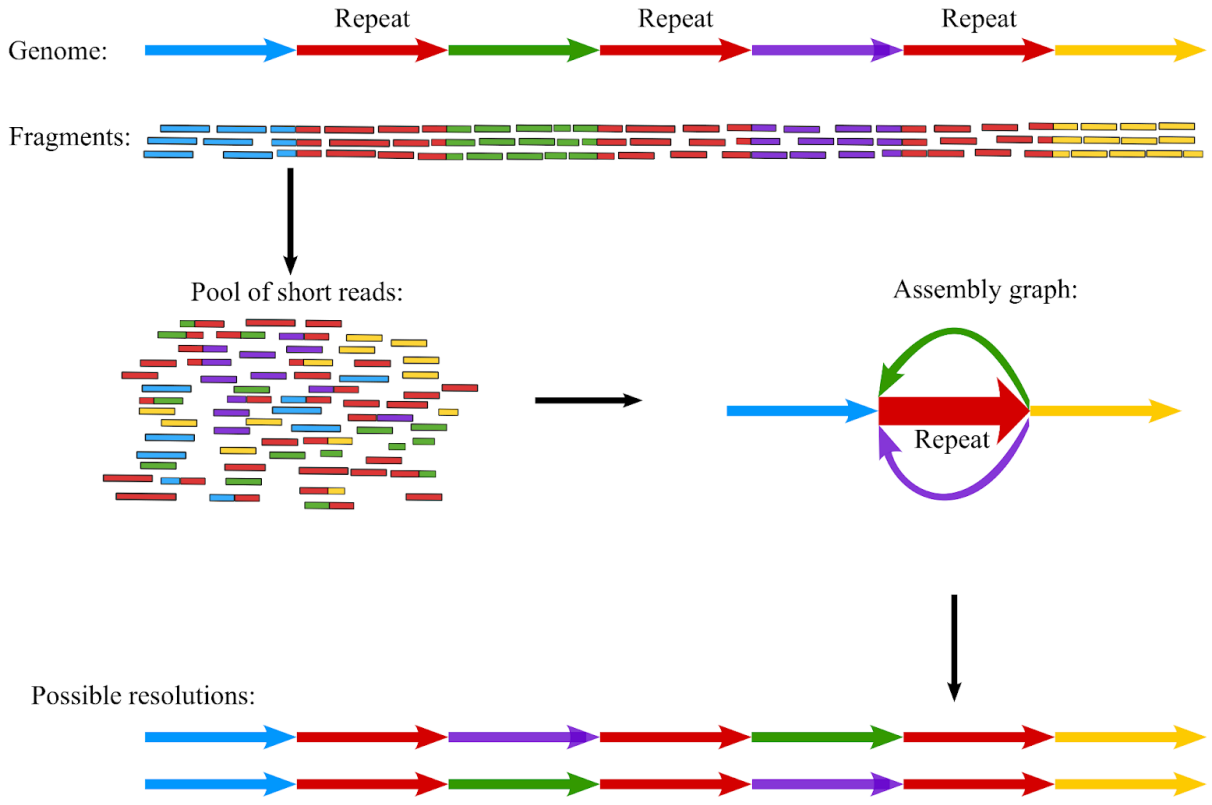


Figure 1.5 | Unresolved short-read assembly near repeats. When read length is shorter than the repeat sequence, ordering of the intervening unique sequences cannot be resolved. Figure taken from Sipos *et al*. 2012 [8].

PacBio sequencing, like Illumina, is still imaging based, relying on addition of fluorescent nucleotides. However, unlike Illumina and similar to ONT, PacBio sequencing is single-molecule sequencing, so no amplification step is required, eliminating the need for the synchronized cycles of Illumina sequencing. Furthermore, a polymerase adds nucleotides in real time as they are imaged; sequencing is closer to an observation of DNA replication than with Illumina sequencing. However, the polymerase used, which is tethered to the bottom of a 'well' where the sequencing takes place, declines in efficiency over the sequencing, limiting read lengths to 30-40kb in length (Figure 1.6a). PacBio sequencing also has a high fidelity method (HiFi), termed

Circular Consensus Sequencing (CCS). This type of sequencing works the same way as PacBio, but a single molecule is circularized, so it can undergo multiple sequencing passes. Adapters on the end used to circularize indicate the start/end of each pass. Read length is further limited by this method, since the ~40 kb length limitation would translate to 4 passes over a 10kb molecule. However, the ability to take the consensus of multiple passes of sequencing eliminates errors, bringing the error rate down from an estimated 8-15% for their single-pass continuous long read (CLR) sequencing to an estimated less than 0.5% for their CCS reads[9,10].

ONT, on the other hand, uses a non-imaging, and non-synthesis based method. DNA is pulled through a 'nanopore' and electrical current is monitored as the DNA moves through the pore. Different combinations of bases, read 5 or 6 at a time (depending on the technology version) have differing currents, so by examining the current in a sliding windowed fashion, by running base-calling software on a profile of the currents, the most likely sequence to have produced those currents can be determined (Figure 1.6b). Basecalling methods have evolved and improved as have the sequencing instruments themselves from initial use of HMMs to higher accuracy methods using deep learning methods to call bases from the electric signals[11].

Error profiles for the various sequencing technologies differ. While Illumina sequencing shows a decline in base quality at the ends of reads, iis prone to more errors (or sequencing failures) in high GC regions, and is more likely to have single-nucleotide errors when an error is present, PacBio and Nanopore data are more prone to indel errors, where multiple bases in a row are erroneously inserted, or bases are erroneously skipped over. ONT sequencing is also prone to

incorrectly reporting the length of homopolymers, as the current remains constant and determining how many bases passed through the pore during that time can be challenging, as the DNA does not move through the pore at a constant rate[11,12].

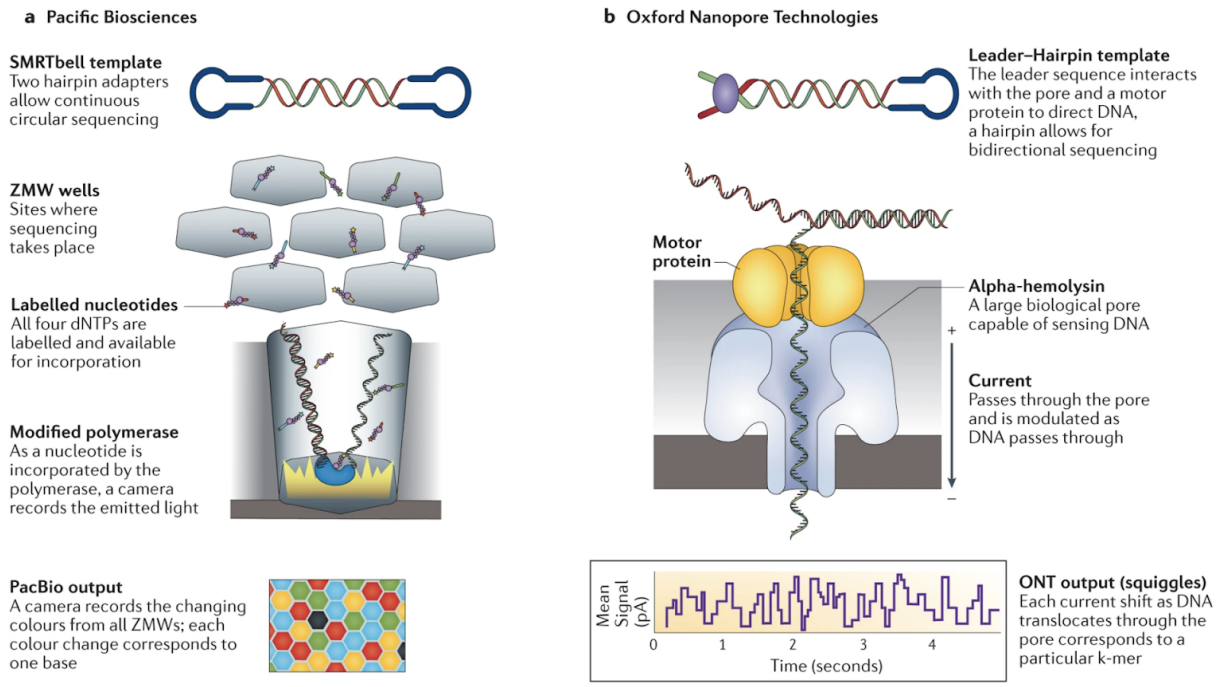


Figure 1.6 | PacBio and Oxford Nanopore sequencing approaches. (a) PacBio sequencing attaches a single molecule to the bottom of a "zero mode waveguide" well, where a polymerase attaches fluorescent nucleotides. This is imaged from above as nucleotides are added; many wells, each with one read, are imaged simultaneously. (b) Nanopore sequencing uses a motor protein to pull a DNA strand through a pore; the current as it passes through the pore is measured, and signal can then be matched to *k*-mers, a process referred to as basecalling. Figure from Goodwin *et al*, 2016 [12].

Long reads provide the benefit not only of being able to span larger repeats in the genome, creating better resolved assemblies, including recently, the first complete telomere-to-telomere assembly of a human chromosome, but also enable the discovery of more variation via read mapping based approaches. With short reads, if an individual has an inserted sequence approaching the size of, or larger than, the read length, the read will not align to the reference genome. However, with a longer read length, the read will align to the reference genome on

either side of the inserted sequence, allowing pinpointing of the insertion within the reference genome. While long reads provided these benefits, Illumina sequencing remains the dominant data type used in whole genome analyses. Long read sequencing costs are declining, but a deep coverage sequencing run is still more expensive with long reads, especially with PacBio technologies. Furthermore, these techniques are lower throughput, and cannot be multiplexed (samples are pooled, then separated). In this thesis, strategies for finding variation using both short and long reads are examined. Though long read sequencing is expected to become more dominant as costs decrease and error rates decline even further, we also discuss strategies for utilizing long reads to gain additional new insights from the wealth of short read data that has already been sequenced. These cross-technology approaches will help advance discoveries while new data is slowly being generated, and maximize the utility of already funded, available, data sets.

1.3 Structural variant detection strategies

Structural variation is typically defined as variants 50 base pairs or larger which differ between individuals of a species. They are often further broken down into the following types: insertions, duplications, deletions, inversions, and translocations (Figure 1.7). As discussed in Chapter 1.1 and 1.2, short reads often struggle to detect large structural variation, particularly insertion sequences approaching or longer than the read length. However, a number of strategies can be used for detecting structural variants via alignment of short reads to a reference genome. Signatures of variation can be detected using coverage, paired end read alignments, or split read alignments. A drop in coverage is an indication of a deletion, whereas duplications would

be indicated by a pile-up of reads, since the reads from each copy of the duplication will align to the same location in the reference (Figure 1.7a). If paired end reads align farther apart or closer together than expected, this indicates a deletion or an insertion, respectively. Inversions can be inferred where paired end reads align in unexpected orientation. Translocations will be indicated by reads aligning to far apart locations in the reference, and duplications might be indicated by unexpected orientation and/or by reads aligning too close together, since a duplication is a form of insertion (Figure 1.7b).

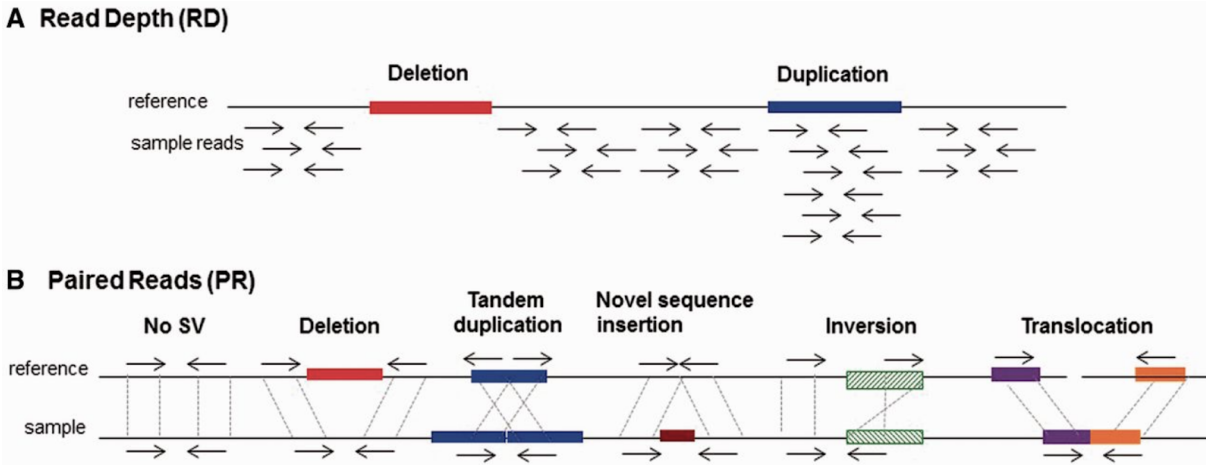


Figure 1.7 | Signatures of short read alignments inferring structural variation. (a) A drop in read depth/coverage can be used to detect a deletion, while a doubling of coverage would indicate a two-copy duplication. (b) Paired end alignments can be used to infer other SV types, based on the alignment distance and orientation of the paired-end reads relative to the expectation based on the fragment size used when sequencing. Figure from Escaramís *et al*, 2015 [13].

Split read alignments, where a read aligns in two distinct pieces rather than contiguously, or alignments with soft clipping (the end of a read does not align at all and is 'clipped' off in the alignment) can also present signatures of variation, though these alignments are not especially common for short reads, as a split alignment of a short read is generally not considered good

enough to report by alignment methods. Spots where the ends of read alignments are all soft clipped might indicate an insertion; the clipped bases being part of the novel non-reference sequence. Split alignments between different parts of the genome would be an indication of a translocation, and split alignments across a small region would indicate a deletion. However, these signature based approaches all infer the presence of a variant; they cannot align a single read through it, and thus the inferred variants may be due to misalignments rather than truly present. Short read variant callers such as Lumpy[14], Delly[15], and Manta[16] focus on analyzing these signatures to call larger variants whereas many callers such as Strelka2[17] or Freebayes[18] only detect small variants (under 50bp). However, these methods are well known to have a high false positive rate[19–21] and typically cannot report any novel sequence at a variant site even if a variant is detected, as reads consisting of novel sequence won't align to the reference genome. As these approaches are known to be error prone, commonly pipelines take the consensus of multiple short read structural variant callers and report the highest confidence variants as those called by multiple tools and signature types[22–24], but even these consensus strategies yield erroneous calls when read alignments systematically produce signatures that don't reflect the true variant, and may miss variants called correctly by only a single caller.

Long read sequencing has dramatically improved structural variant detection. Studies performing structural variant detection from the alignment of long reads to the reference genome have reported on the order of 20,000-30,000 structural variants per healthy individual relative to the reference, far more than previously expected[21,25–27]. Although variant detection is affected by the alignment algorithms used, algorithms which are necessarily different for short

and long reads, long read aligners are able to align through a variant, provided there is sufficient sequence matching on either side of the variant in the read. This allows not only the discovery of new variants with weak signatures from short reads and base pair level resolution of novel inserted sequence, but has also illuminated systematic errors in short read variant calling, such as insertions around repetitive regions being called erroneously as translocations due to mis-alignments of reads in repetitive sequence (Figure 1.8). These systematic errors are particularly problematic in short read variant calling. Since the read alignments show clear signatures which are likely to be called by multiple variant callers, consensus methods will not eliminate these calls.

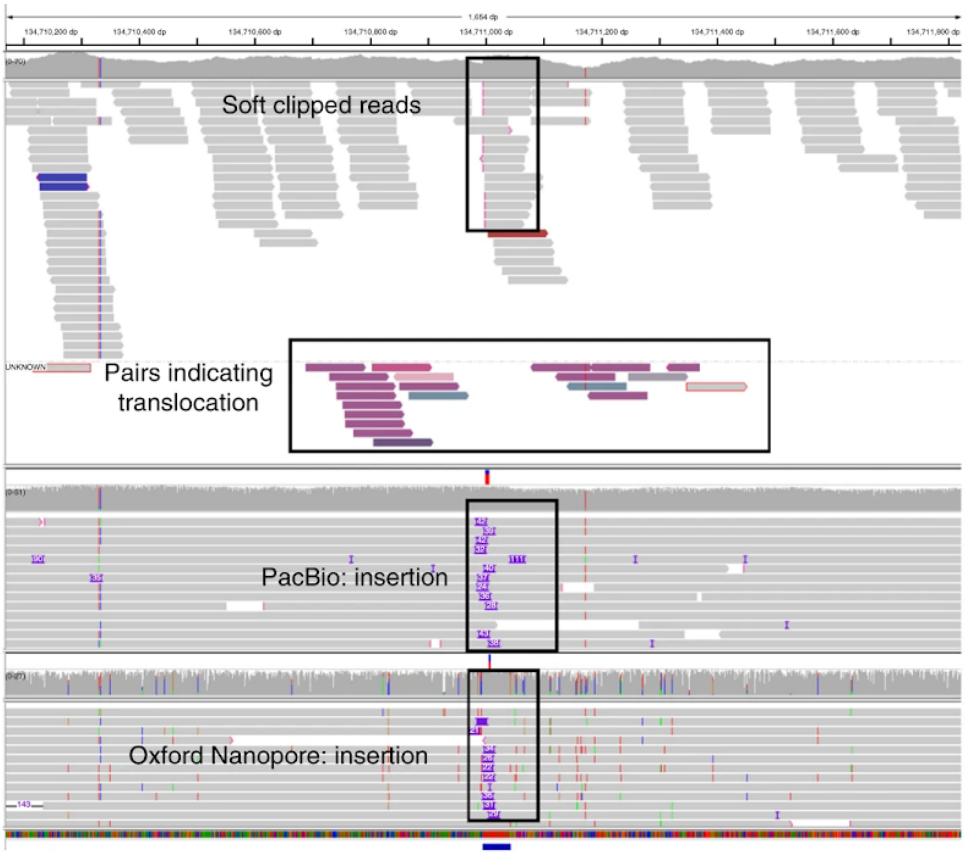


Figure 1.8 | Alignments of short reads and long reads around disagreeing variant calls. Short read alignments lead to a translocation call, as reads are soft clipped, and pairs align to another chromosome. However, both PacBio and ONT long reads indicate an insertion (indicated by the purple bands in the alignment). It is likely the inserted sequence is similar sequence on another chromosome, causing mis-alignments of the short reads and thus a false translocation signature. Figure from Sedlazeck *et al* 2018[26].

However, as long read sequencing data is still limited, examining structural variation on a population level, as short-read studies have done, is only beginning to become a possibility. While Iceland has produced a variant call set on 1,817 individuals, the long read data is not publicly available[28]. Additionally, as Iceland is a fairly homogenous population, this study is unlikely to reveal the scope of human structural variation. Other studies surveying a range of populations with long reads are considerably smaller; to date, there are perhaps dozens of publicly available long read sequenced human samples[27,29]. If we are to capture the scope of human variation, waiting for long read data to be generated that approaches the scale of available short read data will take at best years, and at worst may take a decade or more, particularly as long read technologies are lower throughput. As we continue to produce long read data, utilizing information from a handful of individuals may help enable population screening on short-read datasets. For example, in Figure 1.8, if an algorithm is aware, *a priori*, that some individuals are known to have an insertion at that location, then given only short reads, it would be reasonable to produce an insertion call, rather than a translocation call, since the signature seen in short reads could also be consistent with an insertion, and we have a prior that an insertion is likely. These hybrid approaches may provide new insights, without the need for massive long read data sets; novel methods to perform these analyses, and applications in cancer data sets, are covered in Chapter 4.

1.4 Pan-genomic scale sequencing and approaches for human populations

The ability to compare a newly sequenced individual to a reference and find differences has enabled myriad discoveries and innovations, and in human genomics this ability forms the basis

of thousands of studies seeking the genetic origins of disease. However, as discussed in Chapter 1.3, capturing variation based on alignment to a single reference genome has many limitations, especially when using short read sequencing data. One tactic to better capture the variation missed by using a single reference, is to create and utilize a 'pan-genome', a collection of all the DNA sequences that occur in a species. Ideally, this pan-genome structure, containing all known variation, could then be aligned to, improving read alignment due to missing variation in the reference genome.

Cataloguing the DNA from all individuals in a species is a daunting task. The first pan-genomes were developed for small, easy-to-sequence bacteria, but even in that context, pan-genomes provided novel scientific insights. The consideration of genetic diversity within bacterial species has contributed to our understanding of underlying differences in pathogenicity, virulence and drug resistance, and can even help predict how pathogenic a new strain will be[30–39].
Pan-genome studies of plants and animals remained elusive at first, due to the large genome sizes and vast amounts of intergenic sequence in these species. However, in recent years, thanks to dramatic improvements in the efficiency of sequencing technology, the scientific community has been able to sequence dozens, hundreds, or even thousands of individuals of a single plant or animal species[40]. Additionally, new long-read sequencing technologies now allow us to better assemble repetitive regions of large genomes, including centromeric regions, that are difficult to characterize with short reads[41,42], as described in Chapter 1.1 and 1.2.

Human sequencing too has accelerated. Over the past few years, a flurry of publications have described large collections of newly sequenced human genomes, including population-specific cohorts from Iceland[28,43,44], Denmark[45], Sweden[46], Papua New Guinea[47], Mongolia[48], and Africa[49–51], and large-scale surveys of the entire world[19,52–54]. As these genome collections have accumulated, computational scientists have been working to develop new methods to detect, represent, and analyze large-scale structural variants, which had previously been sidelined while most genetic studies focused on single-nucleotide polymorphisms (SNPs). New representations must be able not only to capture the variation from large collections of genomes, but also to enable efficient means of searching these genomes. Regardless of what methods are chosen, it is now clear that the community must move beyond reliance on a single reference genome. While the use of a single reference has advanced genetics immensely, it has not, as some had hoped, allowed us to find the cause of all genetic disease, a shortcoming that prompted some commentators to call the Human Genome Project a failure[55,56]. Although we now know that many diseases are caused by complex mixtures of multiple genetic variants, if we are to attempt to uncover the genetic causes of many still-unexplained diseases, one of the many factors we must consider is the vast genetic diversity present in the pan-genome.

The concept of a pan-genome was first described by Tettelin *et al*[32] in 2005, in the context of bacteria. They described a pan-genome as a "core genome containing genes present in all strains, and a dispensable genome composed of genes absent from one or more strains and genes that are unique to each strain"; under this definition, the pan-genome captures the whole of the genic content of a species. The dispensable genome is often further sub-divided

into genes unique to one strain (termed 'unique genes') and genes shared between some but not all strains (termed 'accessory genes') (Figure 1.9 a). Restricting the pan-genome to gene content makes less sense in eukaryotes, particularly those with large genomes (>500 Mb) where more than 50% of the genome may be intergenic, and where the gene sequences themselves are dominated by long introns[57]. In addition, eukaryotes do not exchange DNA as freely as bacteria, making their gene content much more stable. For a species such as humans, where exons occupy only ~2% of the genome[58], a pan-genome comprised of only exonic sequences would yield little information about within-species differences. Thus a eukaryotic pan-genome is commonly defined to include all of the DNA sequence in a collection of genomes, not just the genes. Although eukaryotic pan-genome studies sometimes borrow the terms core and dispensable genomes, in eukaryotes these descriptors refer additionally to intergenic sequences, rather than sets of genes, with unique sequences referred to as 'singletons' (Figure 1.9 b).

In the past several years, large-scale human sequencing projects have become increasingly common. No project to date has produced a comprehensive, analyzable, human pan-genome that surveys a wide variety of human populations, captures both genic and intergenic variation, and incorporates this variation into a single utilizable pan-genome. Efforts are underway, however, to create population-specific pan-genomes, as well as to discover as many human SNPs and structural variants as possible, and a recent National Human Genome Research Institute-funded initiative has been launched to build a human pan-genome reference from 350 diverse individuals[59]. With continued development of computational methods capable of

handling larger and larger datasets, these variant catalogues may ultimately provide the data needed to perform pan-genomic analyses in humans.

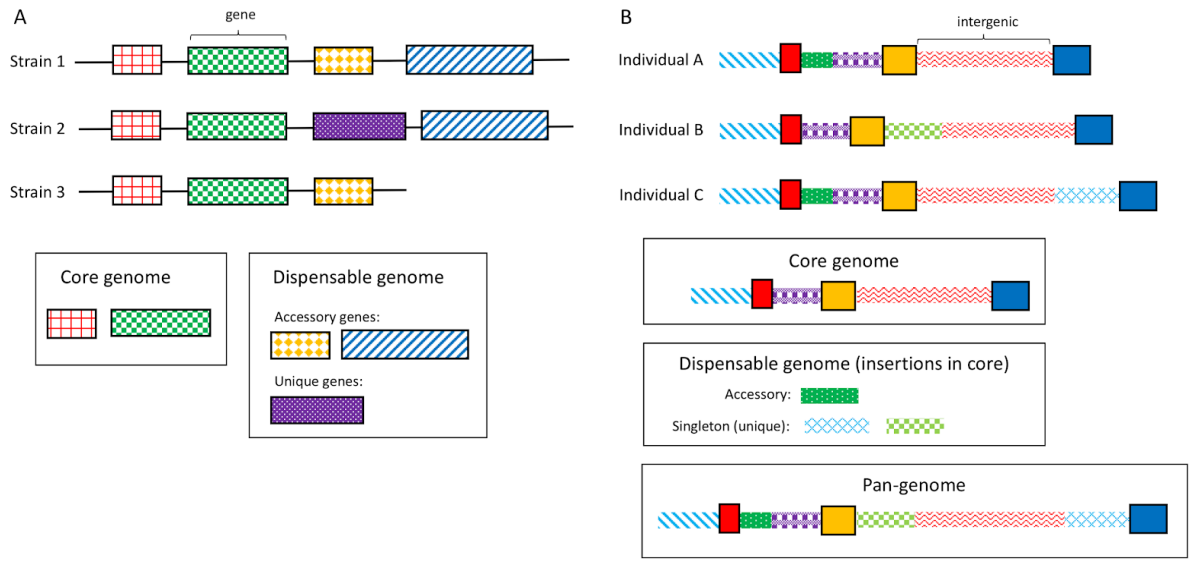


Figure 1.9 | Core and dispensable genomes. **a)** Bacterial and other prokaryotic genomes consist predominantly of genes with little intergenic sequence. The core genome of a species consists of genes shared by all strains. The dispensable genome is made up of genes shared by some but not all strains (accessory genes) and genes present in only one strain (unique genes). Together the core and dispensable genomes make up the pan-genome. **b)** Eukaryotic genomes are not highly variable in their genic content. Pan-genomes consider intergenic sequence as well as genes, resulting in an ordered pan-genome of all sequence present in at least one individual.

Scientists have been cataloguing human variants since well before the completion of the Human Genome Project. However, with the completion of a full reference genome came the ability to catalogue variation genome-wide, leading to the creation of large databases including dbSNP[60] and ClinVar[61], as well as continued updates to pre-existing databases such as Online Mendelian Inheritance in Man (OMIM)[62]. ClinVar and OMIM track variants of clinical interest or with known phenotypic associations, although nearly all of the variants tracked to date are SNPs and small insertions or deletions (indels) relative to the reference genome. While these variants

can be incorporated into genome analyses using SNP-aware aligners such as HISAT2[63], mrsFAST-Ultra[64] or SNPwise[65], we now know that any given individual is likely to contain on the order of 20,000 structural variants (>50 bp) relative to the reference genome[9,21,25,26]. More recent databases such as dbVar, DGVa[66], and DGV[67] aim to catalogue these larger variants, although they cannot yet be easily incorporated into most standard alignment and subsequent analysis pipelines. Several projects have attempted to survey the landscape of human structural variation across the globe, including the 1000 Genomes Project (1KGP)[19], Trans-Omics Precision Medicine (TOPMed)[68], and the Simons Genome Diversity Project[52].

The 1KGP was the first attempt at a large-scale global project for human genome sequencing. The 1KGP was performed in three phases, initially collecting SNP array data and later generating low-coverage (mean 7.4×) whole-genome sequence (WGS) data for 2,504 samples from 26 populations. (In 2019, an updated re-sequencing of these 2,504 genomes was released to improve data quality and consistency. However, to date no studies have been published analyzing this new data release.) An analysis of structural variants in the WGS data reported over 40,000 deletions, 6,000 duplications, nearly 3,000 copy number variants, and nearly 17,000 mobile element insertions by comparison with the human reference genome[69]. 60% of the variants detected were novel relative to the pre-existing Database of Genomic Variants, a database consisting of variants reported from 55 studies at the time of its publication in 2013[67]. In addition to reporting novel variation, one major finding of the 1KGP was a detection of homozygous deletions of large portions of 240 human genes[19]. The discovery that these genes are missing or severely altered in many individuals studied indicates that these genes are part of

the dispensable genic pan-genome, a concept infrequently considered in human genomics. This dispensable gene set was enriched for two classes of proteins, glycoproteins and immunoglobulins, and nearly all of the deletions were found in multiple populations. Other deleted regions in their set of over 40,000 deletions are likely to represent dispensable non-genic regions. Although these findings from the 1KGP are an important step in understanding the dispensable and core human pan-genome, deletion discovery is only one step. The reference genome is missing dispensable sequences as well, which would appear as insertions in the 1KGP samples, but low-coverage WGS data is ill-suited to discovering novel insertions and this was not attempted.

Other global projects have examined novel sequence content. The Simons Genome Diversity Project generated deep coverage (30–40×) in short-read sequencing of 300 individuals from 142 diverse populations[52]. The project assembled sequences that failed to align to the reference genome and discovered 5.8 Mb of novel, non-repeat sequences in the collection. They also catalogued 34.4 million SNPs, 2.1 million small indels, and 1.6 million short tandem repeats. Many of these variants — up to 11% of heterozygous SNP variants in one population — were missing from the 1KGP variant calls, despite the 1KGP dataset containing more individuals, highlighting the need to continue collecting additional samples from diverse populations. A more recent project, TOPMed, has examined short-read WGS data from 53,831 individuals, and using a similar method of assembling unaligned reads discovered 2.2 Mb of novel sequence[68]. Although the TOPMed data contained many more genomes than the Simons Genome Diversity

Project, the investigators discarded any sequence without a good match to one of five hominid genomes, perhaps explaining why they reported a smaller amount of novel sequence.

All of these global projects have limitations. Each of them utilized short-read sequencing data (usually 100-bp reads) that they aligned to the human reference genome, so although some variation can be uncovered, the data necessary to build a pan-genome — that is, the union of all sequences in all humans — remains elusive, in part because reference-based genome assembly methods will entirely miss large insertions in other genomes. Furthermore, none of these large projects has had a primary goal of creating a human pan-genome, and in each case the analysis of novel sequences was secondary to their main findings. Each study has contributed snippets of what is needed, such as dispensable sequences deleted or inserted in many individuals, as well as other detectable variation, small and large, but this variation has not been aggregated in any way into a pan-genome. Although we now have a comprehensive gene set of core and dispensable genes for many bacteria and some plants, even this limited, genic pan-genome view does not yet exist for human populations, in part because we still lack a standardized comprehensive human gene set[70,71]. The pan-genome with intergenic variation included thus remains even more elusive.

In addition to the global projects that have touched on discovering novel sequence insertions, several recent efforts have focused solely on discovering these novel non-reference sequences within and across populations, utilizing both short- and long-read technologies. Many of these efforts state the explicit goal of building a pan-genome, although to date no project has

characterized the insertion and deletion landscape completely enough to generate a full pan-genome, even for a single homogeneous population. Many of the efforts to do so are ongoing, but they remain complicated by the difficulty of determining which repeat sequences are truly novel, and without telomere-to-telomere assemblies of all human chromosomes, it is difficult to tell where repeat copies fall within each individual genome. Thus, definitions of novel sequence vary widely between projects, and as a result so does the amount of novel sequence discovered. Estimates of novel sequence in human populations vary from 0.33 Mb in 15,219 Icelandic individuals, to our estimate of 296.5 Mb in 910 African-ancestry individuals (see Chapter 2). Recent estimates are presented in Table 1.1.

It remains unclear how much of this non-reference sequence is shared across individuals, and thus it is unknown how many individuals must be sequenced before the human pan-genome can be considered complete. We expect, though, that far fewer individuals are needed if only non-repetitive sequence is being considered, since nearly 25 times as much sequence has been found on average in studies considering repeats as opposed to studies which do not (Table 1.1). While non-repetitive sequences may be simpler to analyze, repeat elements can have substantial biological effects on gene expression[72] and disease-related phenotypes[73,74] and these sequences should not be overlooked.

Table 1.1 | Reported novel sequences from efforts to examine structural variation in large cohorts of human individuals.

Population and consortium (if applicable)	Number of individuals	Data type	Total novel sequence reported	Average per individual	Additional requirements	Publication Year	Refs
Swedish, SweGen	1,000 [Subset of 2]	Short read [Long read]	46 Mb [17.3 Mb]	0.6 Mb [12.1 Mb]	Over 300 bp [Over 100 bp]	2019 [2018]	[46] [75]
Han Chinese	275	Short read	29.5 Mb	~5 Mb fully unaligned +	Over 500 bp	2019	[76]

				~6 Mb partially unaligned to reference			
Mixed, TOPMed	53,831	Short read	2.2 Mb	0.2–0.5 Mb	Must align to a hominid genome	2019	[68]
Mixed	154	BioNano maps, linked reads (10X Genomics)	60 Mb	14.2 Mb	>2 kb	2019	[77]
Mixed	15	Long read	21.3 Mb	6.4 Mb	Not in peri-centromeric regions, over 50 bp	2019	[27]
African-ancestry, Consortium on Allergy in African-ancestry Populations	910	Short read	296.5 Mb	2.5 Mb	>1 kb	2019	[78]
Mixed	17	Linked reads (10X Genomics)	2.1 Mb	0.71 Mb	Breakpoint resolved, over 50 bp of non-repetitive content per sequence	2018	[79]
Icelandic	15,219	Short read	0.33 Mb	0.16 Mb	Non-repetitive, breakpoint resolved	2017	[43]
Danish, Danish Genome Project	150	Short read	>15,000 insertions*,‡	Not reported	>50 bp	2017	[45]
Dutch, Genome of the Netherlands	769	Short read	4.3 Mb	Not reported	>150 bp	2016	[72]
Mixed	10,545	Short read	3.26 Mb	0.7 Mb	Non-repetitive, >200 bp	2016	[53]
Mixed, data from 1KGP	45	Short read	61.6 Mb	17,700–20,500 insertions*,§	No size or other restrictions reported	2016	[80]
Mixed, The Simon's Genome Diversity Project	300	Short read	5.8 Mb (13 Mb with repetitive elements)	Not reported	Non-repetitive, >500 bp	2016	[52]
Japanese, Tohoku Medical Megabank Organization	1,070	Short read	9,354 insertions*	45 insertions*	>1 kb	2015	[81]

Summary of reported novel sequences from global and population efforts to sequence and analyze structural variation in large cohorts of human individuals. *Did not report number of bases. ‡ Estimated based on Figure 2b from Maretty et al 2017 [45]. § Estimates separated into average number of contiguous sequences per population with at least a partial match. The 61.6 Mb reported was in 30,879 insertions.

However, cataloging variation is just the first step in creating a human pan-genome. Although databases of SNPs and structural variants (e.g., dbSNP and dbVar) provide a valuable resource for genetic analysis, a comprehensive pan-genome, in whatever form it is stored, is likely to present considerable new challenges for scientists who wish to use it. The amount of sequence data alone is likely to be extremely large, especially if the pan-genome includes all variants of repeat sequences. In addition, the number and variety of rearrangements is both large and difficult to capture in a form that is easy to use with current bioinformatics tools. To date, no computational approach is practically scalable enough to represent and analyze a full human pan-genome created from thousands or millions of individuals.

Currently the most commonly used approach to include divergent human sequences in a genetic analysis is to simply include these extra sequences when performing read alignment to the reference genome. The reference genome already contains several hundred of these alternative or 'alt' sequences, although they do not represent any systematic attempt to capture human variation. This strategy poses a number of problems. First, although the Genome Reference Consortium provides locations for the alt sequences and alignments to the main chromosomes of the reference genome[82,83], most sequence alignment programs were not designed to handle variant information provided in this manner. As a result, most aligners simply treat the alt sequences as additional sequence tacked onto the genome, and as a result the variants are treated as repeats. Some aligners such as bwa[84] have created 'alt-aware' modes to account for this, but even with these fixes, this approach is not sustainable. As we continue to add divergent sequences, storing and searching the reference genome becomes increasingly

space- and time-intensive. Furthermore, including these additional sequences separately does not accurately represent the underlying biology. Although in some cases we know where these sequences belong in the genome, what sequences they are alternatives to, or what populations they are prevalent in, that information is lost by simply including them as additional sequences and continuing to use current algorithms for alignment.

There are a number of approaches for capturing pan-genomic data which allows efficient storage and alignment. These approaches, primarily based on graph based representations of the genome and its variants, pose other challenges, however. These are well described elsewhere; please refer to recent reviews by Sherman & Salzberg[85], The Computational Pan-Genomics Consortium[86] and Paten *et al*[87]. for additional details. Graphical representations tend to provide compact representations for a pan-genome, but the goal is not only to store, but also to analyze pan-genomic data. Given that short-read sequencing (with read lengths in the range 100–250 bp) is the current standard, this means that researchers must be able to align short-read data to the pan-genome representation. Short reads are difficult to align accurately in repetitive regions, even when aligning to a linear reference. Reads may be misaligned if, for example, a SNP or a sequencing error causes them to be identical to a different copy of a repeat elsewhere in the genome. By adding in large numbers of variants, we increase the number of places a repetitive read might align, and increase the chances that a read might be aligned to an incorrect location (Figure 1.10). The study describing the FORGe variant prioritization tool demonstrated that when 8–12% of known SNPs are included, graph aligners such as HISAT2 have the fewest number of incorrectly mapped reads. However, when the

number of variants included is increased beyond that, accuracy declines[88]. Although only SNPs and small indels were examined in that analysis, the logic extends to structural variants as well, particularly when variants belong to a high-copy-number repeat class, such as the HSAT II and III centromeric repeats. Another study demonstrated that whereas graph-based mapping with vg[89] and SevenBridges[90] yields higher accuracy than linear alignment on reads that contain known variants, linear genome alignment is superior when the reads do not contain variants[91].

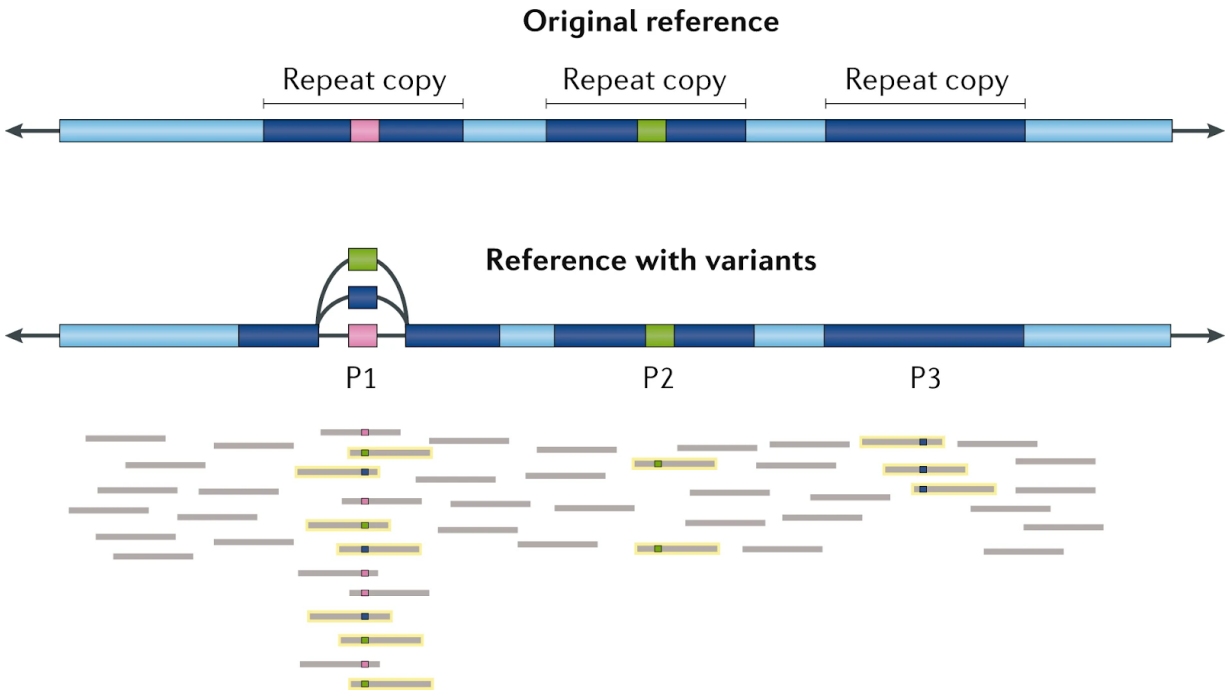


Figure 1.10 | Inclusion of variants complicates read alignment. A graph-based representation includes alternate variants (blue, green) at position P1, whereas the reference contains only the pink reference allele. These variants are within a repeat (dark blue). The addition of each alternate variant increases alignment ambiguity. The six reads with the blue variant allele align perfectly only to P3 in the original reference, and now align to P1 and P3 equally well. Likewise, the six reads with the green variant allele now align to P1 or P2 perfectly, not just P2. Ambiguous reads are highlighted with yellow outlines.

Despite these challenges, a clear advantage of human pan-genome analyses is that scientists can discover variants that are missing from the reference genome and then link those variants

to phenotypes, which might include both beneficial and harmful traits. For example, any sequence longer than a few hundred nucleotides that is missing from GRCh38 is essentially invisible to most downstream analysis tools, regardless of how many individuals are sequenced, because any reads containing that sequence will simply fail to align. If such a sequence contains a disease-causing or disease-preventing variant, that variant will be undiscoverable unless this sequence is included in the analysis (Figure 1.11). An illustration of this arose in the Icelandic human sequencing project, where their examination of 15,219 Icelandic individuals found a 766 bp insertion at high allele frequency (65%), the presence of which was found to significantly correlate with a decreased risk of myocardial infarction[43]. Another recent study discovered a

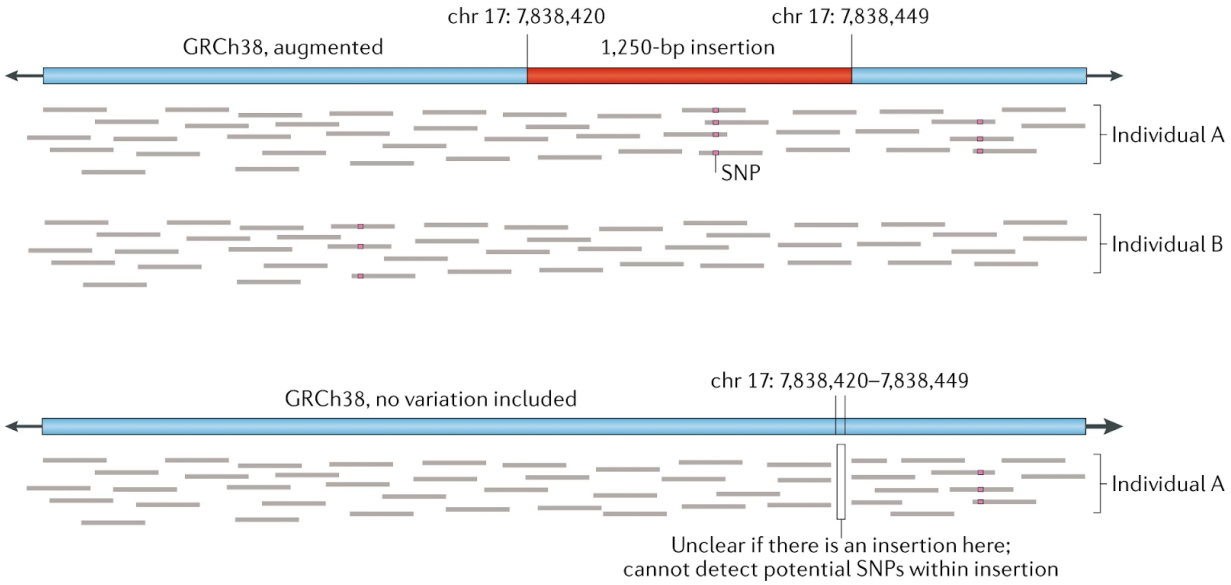


Figure 1.11 | Augmenting the reference genome can lead to novel variant discoveries. When the reference genome sequence is augmented with a known insertion, reads will align to this region for individuals containing this insertion. The 1,250-bp insertion included on chromosome 17 (chr 17) is within the gene *KDM6B* and has been reported in numerous studies[27,43,68,85] including at a frequency of 1 in the Trans-Omics Precision Medicine (TOPMed) dataset of over 53,000 individuals[68], and thus appears to be present in all or most individuals. With the insertion included in a pan-genome reference, reads from sequenced individuals will align to the region, allowing for the detection of single-nucleotide polymorphisms (SNPs). Here a SNP can be detected that is present in individual A but not individual B. However, when no pan-genomic variation is included in the reference, neither the insertion sequence nor the SNP in individual A can be detected. The depicted coordinates and the length of the *KDM6B* insertion were taken from Sherman et al 2019[85], although they are nearly identical in all reports.

repeat expansion causing neuronal intranuclear inclusion disease, a fatal neurodegenerative disease that causes symptoms ranging from deterioration of motor function to dementia[92]. That study utilized long reads to discover the repeat expansion, and then demonstrated that the repeat expansion could be genotyped in other individuals using short reads alone. Although our ability to efficiently sequence and analyze large collections of human genomes is still limited, these recent examples are a demonstration of the potential to detect new and important variants as we become increasingly able to analyze human pan-genomes. We present several examples of our work detecting these variants, and attempting to determine their relevance in both population genomics and in clinical settings in the remainder of this thesis.

Chapter 2: Novel sequence discovery in 910 genomes of African-descent

A version of Chapter 2 has been previously published as:

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., ... & Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, *51*(1), 30-35.

2.1 Background: Including genomic diversity in human sequencing

Since its initial publication[2,3], the human genome sequence has undergone continual improvements aimed at filling gaps and correcting errors. The latest release, GRCh38, spans 3.1 gigabases (Gb) with just 875 remaining gaps[4]. The ongoing effort to improve the human reference genome, led by the Genome Reference Consortium, has in recent years added alternate loci for genomic regions where variation cannot be captured by single nucleotide polymorphisms (SNPs) or small insertions and deletions (indels). These alternate loci, which comprise 261 scaffolds in GRCh38, capture a small amount of population variation and improve read mapping for some data sets.

Despite these efforts, the current human reference genome derives primarily from a single individual[5], limiting its usefulness for genetic studies, especially among admixed populations such as those representing the African diaspora. In recent years, a growing number of researchers have emphasized the importance of capturing and representing sequence data from more diverse populations and incorporating these data into the reference genome and genomics studies in general[93–95]. The alternate loci in GRCh38 offer one possible way to add such diversity, although it is unclear whether such a solution is sustainable as more populations are

sequenced. Among other problems, the addition of alternate loci as separate contigs can mislead sequence alignment programs, which were designed under the assumptions that each read has a single true point of origin, and that the genome is represented as a linear haploid sequence[83].

The lack of diversity in the reference genome poses many challenges when analyzing individuals whose genetic background does not match the reference. This problem may be addressed by using large databases of known SNPs (e.g., dbSNP[60]), but this solution only addresses single-base differences and small indels, and is not adequate for larger variants. Findings from the 1000 Genomes Project indicate differences between populations are quite large; examining 26 populations across five continents, 86% of discovered variants were found to be present in only one continental group. In that study, the five African populations examined had the highest number of variant sites compared to the remaining 21 populations[54].

One way to address the limitations of a single reference genome is to sequence and assemble reference genomes for other human sub-populations. The 1000 Genomes Project, Genome in a Bottle, and other projects have assembled draft genomes from various populations, including Chinese, Korean, and Ashkenazi individuals[96–101]. Others have used highly homogenous populations (e.g., Danish, Dutch, or Icelandic individuals) together with assembly-based approaches to discover SNPs and structural variants (SVs), including up to several megabases of non-reference sequence common to these populations[43,45,72,102], as described in Chapter 1.4 and Table 1.1. While these variant analyses are a step in the right direction, to date none have

produced a reference-quality genome that can replace GRCh38[4], although this is an explicit goal of the Danish Genome Project (http://www.genomedenmark.dk/english/about/referencegenome/).

While efforts to produce new reference genomes are worthwhile, attempts to create a "pan-genome" of a human population; i.e., a collection of sequences representing all the DNA in that population, are rare. Although pan-genomes are often created for bacterial species[32,103,104], pan-genomes for animal[105,106] or plant species[107–112] are still in their infancy compared to these bacterial pan-genomes. These eukaryotic pan-genomes often consist of variant calls on a reference genome rather than a comprehensive pan-genome including novel sequences, though some recent efforts attempt to capture novel sequence from *de novo* assembly as well, including the goat pan-genome, which captured 38Mb of novel sequence relative to the goat reference genome[106]. For more details on pan-genomics and variant calling in general, refer to Chapter 1. The lack of pan-genomes is due in part to the technical challenges of assembling many deeply-sequenced genomes *de novo* and combining them into one genome. While the Danish Genome Project focused on 50 trios of non-admixed individuals (removing admixed samples from their study[45]), our study focuses on a highly heterogeneous group of admixed individuals. Because the human reference genome is largely complete (i.e., the sequence has very few gaps), our strategy for creating a pan-genome focused on finding large insertions. This approach, although computationally demanding, made the African pan-genome assembly process described here feasible.

A 2010 study that sequenced one Asian and one African individual used the novel sequences identified to estimate that a full human pan-genome would contain an additional 19-40 megabases (Mb) not in the current reference genome[96]. Recent efforts to sequence a Dutch population and a set of 10,000 individuals have supported this estimate, reporting 4.3 and 3.3 Mb of non-reference sequences respectively[53,72], however neither study was designed with the primary goal of discovering long, non-reference sequences. A 2017 study, where two haploid human genomes (hydatidiform moles) were sequenced using long reads, estimated that a single diploid genome may differ by as much as 16 Mb from the reference genome[20]. As we describe here, our analysis of 910 deeply sequenced individuals, all from the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA)[51], produced a much larger amount of novel sequence (i.e. sequence absent from GRCh38) in the African pan-genome spanning 296.5 Mb. This finding, which we published in 2019 in Nature Genetics, has since seen validation in the form of other studies discovering large amounts of novel sequence using long read technologies, most notably Audano *et al* [27] which found more than double the amount of average novel sequence per individual, though the study size was much smaller. Our findings here and other recent and ongoing studies demonstrate just how much we are missing when we ignore human diversity and base all genomic studies on a single imperfect reference genome.

2.2 Integrating alignment and assembly to discover novel sequence from short-reads

Due to the challenges of *de novo* assembly of short-read data, particularly across a large data set of nearly 1000 WGS sequenced individuals, we present here an approach combining

alignment to the reference genome, and subsequent assembly of only reads which do not aligned to the reference, GRCh38. This strategy allows for the assembly of wholly novel sequences relative to GRCh38, while limiting the computational resources and time needed to perform *de novo* assembly, since assembly is only performed on a small subset of the reads (typically 1-3% of total reads do not align to the reference). This combined alignment-assembly approach will assemble contigs of novel sequence, but the primary caveat is that these novel sequences will not be localized relative to the reference genome.

Attempting to place the novel sequences within the reference genome with this approach presents a challenge, but mate pair information from Illumina sequencing data can be used to try and localize sequences. In cases where one mate was assembled into a novel sequence contig but the mate pair aligned to the reference genome, contigs can be linked to their approximate reference genome location. Utilizing this linking mate information, we attempt to place the assembled sequences into the reference genome, taking care to only report an exact location when it can be unambiguously determined. Additionally, the presence of a contig in many individuals can be leveraged here -- if a contig is present in many individuals in a cohort, it is more likely that mate linking information will be present and unambiguous in at least one individual, allowing for placement. Thus, common sequences within the cohort are more likely to be localized in the reference. As we attempt to build a pan-genome, ensuring the inclusion of common sequences is more critical than including rare, singleton variants, so the examination of large cohorts helps overcome the caveat of placement difficulty using this alignment-assembly hybrid strategy. In future studies, long reads, which are able to span novel

sequence insertions, will not only enable discovery of new novel insertion sequences but may also aid with placement of sequences such as those we report in this study.

We present in the remainder of Chapter 2.2 a detailed methodology for this alignment-assembly- placement strategy which we performed on WGS data from 910 genomes of African ancestry.

Data Overview

We used whole-genome shotgun sequence data from 910 individuals whose genomes were sequenced as part of the CAAPA project, available from dbGaP as accession phs001123.v1.p1. The total data set contains 1.19 trillion (1.19×10^{12}) 100 bp paired end reads, from ~300bp fragments, representing an average of 30-40X coverage for each individual's genome. Sequencing was performed on an Illumina HiSeq 2000. The subjects in the study were all of African ancestry and were selected from 19 populations across the Americas, the Caribbean, and continental Africa (Table 2.1)[51]. Due to the short-read nature of the data, we describe below a hybrid alignment/assembly approach to assemble large non-reference insertions, while avoiding whole-genome assembly, which would use extensive computational resources for poor to mediocre assemblies due to only having 100bp reads at 30x coverage.

Table 2.1 | **Cohorts of CAAPA samples.**

Cohort	Number of Samples
African American (Atlanta)	50
African American (Baltimore-DC)	50
African American (Chicago)	50
African American (Detroit)	50
African American (Jackson, MS)	50
African American (Nashville)	48
African American (NYC)	48
African American (San Francisco)	50
African American (Winston-Salem)	50
Barbados	49
Brazil	47
Colombia	50
Dominican Republic	47
Gabon	34
Honduras	50
Jamaica	50
Palenque	34
Nigeria	50
Puerto Rico	53

Data was collected from 19 distinct cohorts across the Americas, the Caribbean, and Africa resulting in 910 analyzed samples.

Assembly of novel contigs

For each sample, we aligned all reads to GRCh38.p0 using Bowtie2[113], and extracted unaligned reads and their mates using Samtools[114] (Figure 2.1). GRCh38 alternate loci were excluded from the reference index, but were considered later in the process. We then assembled all unaligned reads with the MaSuRCA assembler[115]; if neither mate in a pair aligned to GRCh38, MaSuRCA treated the reads as paired-ends with a fragment size of 300 bp, and if only one mate was unaligned MaSuRCA treated it as an unpaired read.

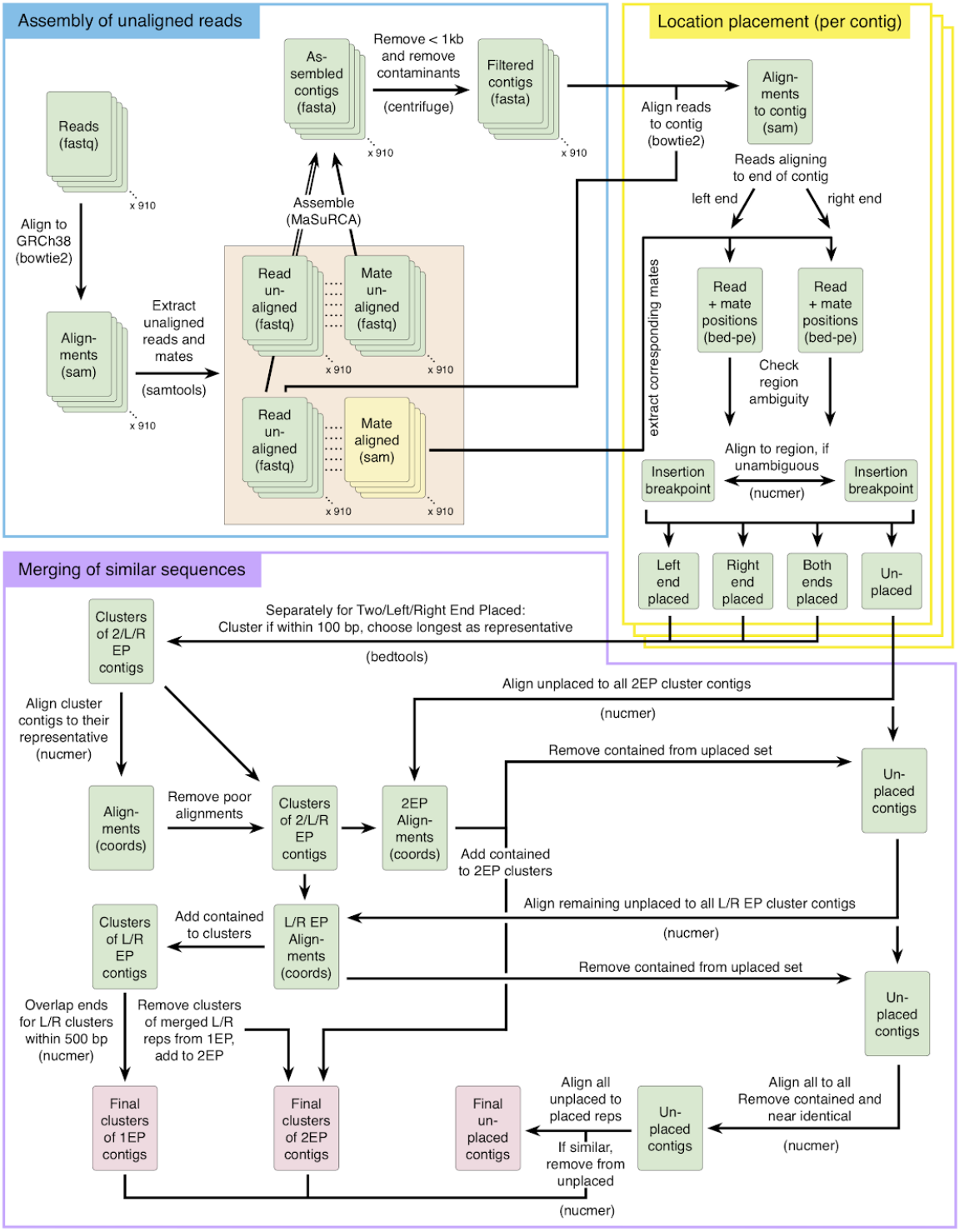


Figure 2.1 | Overview of methods. Raw reads are aligned to GRCh38 and unaligned reads assembled with MaSuRCA. Assembled contigs are then filtered for contaminants with Centrifuge and contigs shorter than 1 kb are removed (blue box). Assembled contigs are placed based on their mate's alignment locations when possible, by checking if over 95% of mates align to the same location. If such a placement is found, the exact breakpoint is determined via a nucmer alignment to the region for each end of the contig (yellow box). Contig placement locations are then compared between all individuals, nearby placements are clustered, and a representative is chosen. All contigs are then aligned to the representatives to determine which samples contain a given placed insertion. Contigs in or aligning to placed clusters are removed from the unplaced set, and the remaining unplaced contigs are aligned to one another with nucmer to remove redundancy and result in a final nonredundant unplaced set of contigs (purple box).

We filtered the resulting assemblies to exclude contigs shorter than 1000 bp (Figure 2.1) and evaluated all remaining contigs with the Centrifuge metagenomics program[116], scanning against the comprehensive NCBI nucleotide database to obtain a taxonomic classification of each contig. We considered any contigs labeled by Centrifuge as non-chordates (e.g., bacterial or viral contigs) to be contaminants and removed them from further consideration.

Positioning contigs within GRCh38

We attempted to place the assembled contigs in a precise location in the human genome using mapping information from paired reads ("mates"). We masked contigs with RepeatMasker[117] with the low complexity option off (-nolow), and used Bowtie2 to re-align all unaligned reads from read pairs in which only one mate had aligned originally. For each read *R* aligning within 500 bases of the end of a contig, we examined the alignment of *R*'s mate to GRCh38 to determine if the contig had a unique placement in the reference genome. The fragment length for all paired-end libraries was 300 bp; by considering reads within 500 bp of the end of a contig we reduced the likelihood that one or both of the alignments was a spurious match. This process resulted in a pool of linking mates corresponding to the beginning and end of each contig.

We then separated contigs into several groups based on their linking information:

1. No linking mates existed on either end of the contig; i.e., the reads mates did not align to GRCh38.

2. Placement was unambiguous (or unique) for at least one end of the contig. We define "chromosome unambiguous" to mean >95% of the linking mates linked to the same chromosome. We define "region unambiguous" to mean that of the >95% of mates aligned to the same chromosome, all mates aligned within 2 kb of each other. When both conditions hold, we say placement is unambiguous. These contigs were further divided into two subgroups:
    a. Both ends of the contig were placed unambiguously, or
    b. Only one end was placed unambiguously.
3. At least one end of the contig was chromosome unambiguous, but neither end was region unambiguous.
4. Neither end was chromosome unambiguous.

For all contigs in the second group, we used NUCmer[118] to align them to the region determined by the linking mates (Figure 2.1). If a contig end had one or more consistent exact matches of at least 15 bases (and no inconsistent alignments), we then determined the contig end's exact insertion location based on alignment coordinates. We permitted an exact two-ended placement only if both ends aligned to the same reference region with the same orientation. The insertion position was either a single breakpoint, if both ends of the contig were placed identically, or a range if the insertion location of the two ends was not identical. For contigs with only a single end exactly placed, we recorded their exact single-end insertion position and the number of overlapping bases (i.e. bases to be trimmed off the end of the contig). One a region was unambiguously determined for a contig end, we performed NUCmer alignments to

determine the exact placement location. For unambiguous contig ends, we aligned the terminal 200 bp of sequence to the region determined by the mate placements. These regions were up to 2 kb in length, which we padded with an additional 500 bp taken from both sides of the region. Alignments were performed without repeat-masking because the region had already been unambiguously identified. We used the parameters `--minmatch 15 --breaklen 1` to disallow gaps or mismatches in the alignments and left all other parameters as defaults. If we found at least one exact match of at least 15 base pairs within 5 bases of the contig end, and all exact matches were consistent with one another, an exact breakpoint was determined by chaining the alignments. The resultant aligned portion of the contig was recorded (to be trimmed off later in the pipeline) and the endpoint of the alignment was recorded as the insertion location for that end of the contig.

Insertion discovery with PopIns

To supplement the list of placed contigs determined by the procedure above, we ran the PopIns program[102], which was used previously for a set of genomes from Icelandic individuals, and was designed to find insertions from a relatively genetically homogenous population. We ran PopIns beginning with the `popins merge` step, using the cleaned MaSuRCA contig assemblies described above. We ran subsequent PopIns steps as recommended in the PopIns documentation, through the `popins place-finish` step. In the `popins merge` step, PopIns produces new contigs by merging those provided to it into new merged contigs. To obtain clusters of placed contigs which could be more easily merged with our pipeline, we aligned all MaSuRCA contigs to each merged contig created and placed by PopIns. We grouped

contigs that fully aligned with over 98% identity into a single cluster representing one insertion and location. For all contigs in each cluster, we then aligned each contig's ends to the placement location with NUCmer, as described above, to attempt to determine its exact placement.

Notably, PopIns generates placements of a single end of a contig using the VCFv4.2 breakend format to specify how much of a merged contig is inserted at a breakpoint. Thus in many cases PopIns output several placements for the same contig that did not agree in orientation or placement location. If we could not verify a placement made by PopIns via our independent alignment of the contig ends to the placement region, we excluded it from the final set of placed insertions. Of the contigs in a cluster that could be exactly placed with NUCmer, if one or multiple contigs had both ends placed, the longest of these was reported as the representative of the cluster and the insertion was added to the set of two-end placed insertions.

If no single contig had both ends placed by NUCmer, up to two representatives were chosen, with the longest contig being chosen as the representative for each end of the contig. This resulted in the potential for two separate one-ended clusters, which were then added to the one-end placed insertion set as follows. All PopIns representatives where the contig had already been placed in a two-ended placement cluster were excluded regardless of location conflicts, as the two-ended clusters necessarily had more evidence supporting them. We further excluded as redundant any PopIns placements within 100 bases of an existing one-ended or two-ended placement location.

Once PopIns placements were incorporated into the one and two end placement sets and clusters had been finalized, we attempted to verify contig placements produced by PopIns. To verify the placements we examined the placement locations of linking mates from all contigs in the cluster of a PopIns placement. Clusters in which fewer than 25% of all linking mates aligned within 5 Kb of the GRCh38 placement location were removed from the placed set. If no mates existed, the cluster was not removed. This resulted in the removal of a number of PopIns placements, including several for which we had determined a one-ended placement with very strong mate-pair support but a PopIns' two-ended placement disagreed.

Clustering of placed contigs

Once contig locations were determined for each individual sample, we aligned all insertions to one another and clustered them to determine which contigs represented the same insertion across individuals (Figure 2.1).

Clustering two-ended placements

For contigs with both ends placed, we ran BEDTools merge[119] to group contigs placed at approximately the same location. We used the -d option with a distance of 10, to allow placements within 10 bases of each other to be combined. We also ran the merge using `-d 100`, which produced identical results. For each resulting region and contig cluster, we chose the longest contig in the cluster as the cluster's representative (*R*), and these representatives formed the initial set of two-ended placed contigs, 2EP. Two-ended placement clusters from PopIns were then added to 2EP. We verified clusters by aligning all contigs in each cluster to its

representative, *R*, with default nucmer parameters and removing from the cluster any contigs that did not have any alignments to *R*. To find the complete set of samples containing each insertion, we then aligned all remaining contigs (including unplaced contigs) to the contigs in the clusters. Any contig aligning with greater than 99% identity that was fully contained within a contig in a cluster *C* and covered at least 80% of the contig in *C* was included in *C* as part of the final set. Contained, 99-100% identical contigs aligning with less than 80% coverage were also included if they had at least 5 linking mates and at least 25% of those mates linked to within 5kb of the placement location. The longest representative contig in each cluster was used as the final insertion sequence for the African Pan-Genome (APG) contig collection (Supplementary Tables 1-2 in Sherman *et al* 2019[78]).

Clustering one-ended placements

We separated contigs with only one end placed into two groups: (1) contigs where the "left" end aligned to the reference, so that the contig extends into a gap to the right of the placement location; and (2) contigs with their "right" end placed, so the contig extends into a gap to the left of the placement location (Figure 2.1). Left and right were determined by the orientation of the chromosomes in GRCh38. We then created clusters separately for the two groups using BEDTools merge (`-d 100`) as described above, identifying the longest representative *R* for each group. This formed the initial set of one-ended placed contigs, 1EP. Any placements within 100 bases of a two-ended cluster (in the set 2EP) were then removed from 1EP, and each contig in these 1EP clusters was aligned to the representative of the 2EP cluster(s) within 100 bases. If

any 1EP contig in the cluster aligned with ≥ 80% coverage and ≥ 90% identity to the 2EP contig, the 1EP contig was added to the 2EP cluster.

We then added PopIns one-ended placement clusters to the right and left placements in 1EP. Then for all clusters, we used NUCmer with default parameters to align contigs within each cluster to the representative *R*. If no alignment was found between a contig and *R*, the contig was removed from the cluster. We then re-aligned all other contigs to those in each of these filtered clusters, excluding contigs already determined to be part of a two-ended insertion. Contigs > 99% identical over their whole length to any member of a cluster *C* and covering at least 80% of the contig in *C* were added to *C*. Contained, 99-100% identical contigs aligning with less than 80% coverage were also included if they had at least 5 linking mates and at least 25% of those mates linked to within 5 kb of the placement location.

We then evaluated the one-ended placements to determine if two contigs might belong to the same longer insertion, where one contig would "fill" the left side of a gap and the other would fill the right side, possibly meeting in the middle. In some of these cases, the contigs might overlap, allowing us to merge them and create a single, longer insertion sequence. If placement positions were within 500 bases of one another, we ran `nucmer --maxmatch --nosimplify`, followed by `show-coords -o` (with annotation) to align the representative contigs of clusters that were candidates for merging. If NUCmer annotated the representatives as identical, or if it found that either contig contained the other with at least 97% identity, we merged the clusters and reported the longer representative contig. In cases

where the ends of two contigs overlapped in the correct arrangement and orientation relative to their placements, we merged the overlapping ends by extending the sequence of the longer contig with that of the shorter contig as indicated by the alignments. The resulting merged sequence and cluster was then moved to the 2EP set. If NUCmer identified other alignments between representatives covering at least 50% of one of the representatives and the clusters shared any contigs (i.e. at least one contig was contained in both representatives), the clusters were merged. However, because these representatives were more divergent the representatives were not merged and the longer representative was reported in the 1EP set (Supplementary Table 1 in Sherman *et al* 2019[78]).

Finally, to remove any potential redundancy from placed clusters, we aligned all representatives from both one- and two-end placed clusters to one another (using `nucmer --maxmatch --nosimplify`) regardless of placement distance. If two representatives aligned with ≥ 98% identity, covering ≥ 95% of one of the contigs, and were placed within 5 kb of one another, these clusters were merged. To determine the representative (and therefore reported placement) of the merged clusters, two-ended placed representatives were favored over one-ended ones, then our placements were preferred over PopIns, then longer contigs were favored over shorter contigs. By merging only placements within 5 Kb, we avoided merging contigs that were similar solely due to repetitive sequences but were unambiguously linked to different locations.

Unplaced contigs

For all unplaced contigs, we ran `nucmer --maxmatch --nosimplify` with a minimum seed length of 31 (`-l 31`) and a minimum cluster size of 100 (`-c 100`) to align all contigs against one another. Contigs contained within another contig and aligning with > 95% identity were removed, and if contigs were annotated as identical by `show-coords` with > 97% identity, the smaller of the two was removed. If the ends of two contigs overlapped by at least 100 bases and a third contig was contained within the joined contigs, the contained contig was also removed. Trimming of up to 100 bases was permitted for finding overlaps. Finally, we aligned all resulting unplaced contigs to the placed representatives pre-trimming. If an unplaced contig aligned with ≥ 80% coverage and ≥ 90% identity, it was removed from the unplaced set, though it was not added into the placed cluster, as it did not meet the stricter placement or containment criteria used to create the clusters.

In an additional attempt to place more contigs in the reference genome, we repeated the placement procedure described above, this time considering only the subset of linking mates that mapped to GRCh38 with a mapping quality >10, and only attempting to place a contig if the contig end had a minimum of 5 such linking mates. This mapping quality criterion decreased the overall ambiguity of the putative locations for unplaced contigs (Figure 2.2), however this additional placement effort only placed 150 additional contigs. We produced a file of putative linking locations for unplaced contigs by examining separately for each end, the linking mates with a mapping quality >10. If greater than 50% of these high-quality linking mates for a given end pointed to the same region, where a region was defined by grouping mates within 2kb of

each other, we reported that region as the putative placement location for that end of the contig, as well as the total number of high-quality mates, and the percentage of those mates linking to that location. For this report, the two contig ends were allowed to putatively link to different locations; in such cases both the start and end regions identified are provided as these are the two most likely placement regions for the contig (Supplementary Table 3 in Sherman *et al* 2019[78]). The putative locations include high-copy repetitive sequences that may be underrepresented in GRCh38, and thus are overrepresented as linking locations.

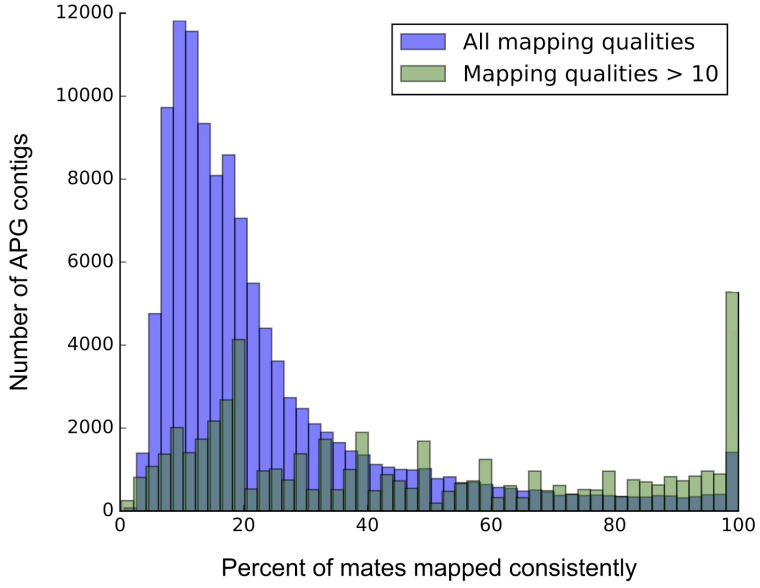


Figure 2.2 | Ambiguity of linking mate placement locations. Mates linking a contig end to GRCh38 were considered for all unplaced contigs. Contigs which had at least 5 linking mates were included in the histogram. The x-axis represents the percentage of mates mapping to the "best" placement region. For example, if a contig had 10 linking mates, and 4 mapped to chr1:50000-55000, 3 mapped to a chr2:88000-90000 and the remaining 3 mates all mapped to different locations, then the percent mapped consistently to the best location would be 40%. When mates of all mapping qualities are considered, there is considerably more ambiguity, although there are also more APG contigs with at least 5 linking mates considered. Although the number of APG contigs with at least 5 linking mates drops from 117,454 (blue) to 53,205 (green) when filtered for mapping qualities above 10, far more contigs can be linked to a placement location with high confidence and the number of contigs where all mates point to the same region roughly triples.

We examined these over-represented genomic locations to which >100 contigs linked in five randomly selected samples (Figure 2.3). All regions in all five individuals had excessively deep coverage, ranging from ~10 times to ~3000 times greater coverage than the 30-40X expected for

these datasets. We also observed that all regions had an abundance of mismatches in the alignments. These mismatches in the read pile-ups indicate that these sequences occur in many copies throughout the genome, with somewhat diverged sequence, but are presented by only a few copies in the GRCh38 reference genome. Many of these sequences occur in or near centromeric regions, in which this phenomenon has been previously detected[37]. Since the contigs linking to these regions did not meet our merging criteria, and the locations linked to are not high confidence, the contigs were left as separate sequences in our APG set.

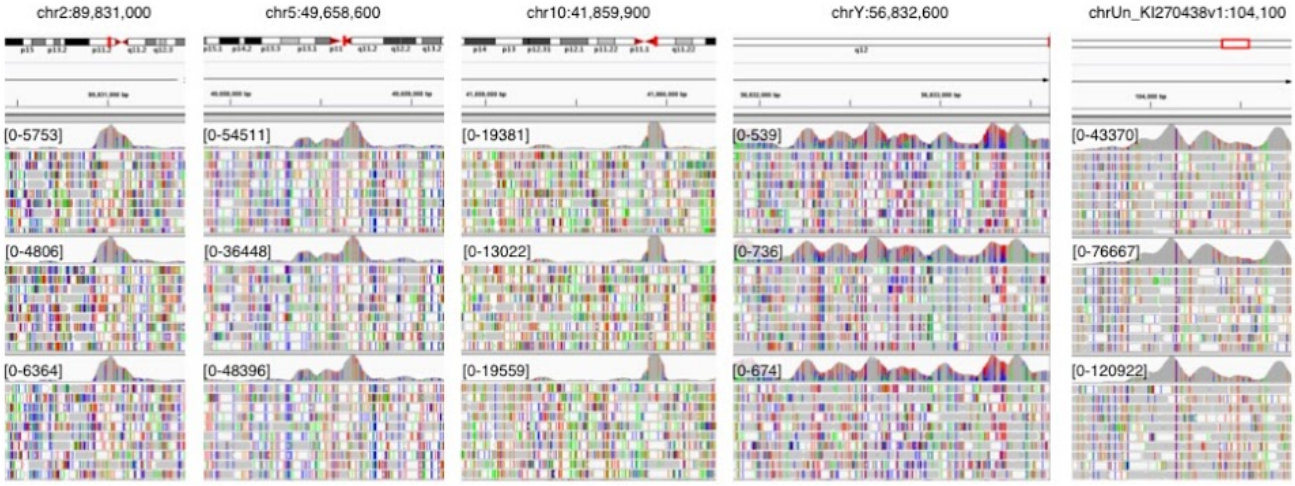


Figure 2.3 | Regions with an over-representation of linking mates. Several genomic regions have an over abundance of APG contigs tentatively linked to them via mate alignment information, though the linkages were not unambiguous enough to meet placement criteria (Supplementary Table 3 in Sherman *et al* 2019[78]). Three randomly selected CAAPA samples all have read alignments to these regions at far greater than expected coverage, ranging from ~15 times (chrY) to ~3000 times (chrUn) greater coverage than the expected (30-40X). Vertical panels are labeled with approximate location of the coverage peak. For each of the three samples the coverage is displayed in the upper frame with the max coverage indicated, and a subset of the aligned reads are shown in the lower frame, with mismatches colored according to the base. All regions have an abundance of mismatches in the alignments. These mis-alignments resulting in read pile-ups may indicate that these sequences occur in many copies throughout the genome, with somewhat diverged sequence, but only occur a single time in the reference genome. In fact, several of these sequences, including those pictured on chr5 and chrUn_KI270438v1, have only been present in the reference genome since the release of GRCh38, and may still be underrepresented in GRCh38 due to difficulties in assembling these sequences. Coverage and alignment images were produced by IGV (Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, *14*(2), 178-192, 2013).

Another consideration in assessing the accuracy of the putative linking mates is repetitive genomic regions, even if the regions are not at deeper than expected coverage, as repetitive sequences are expected, since GRCh38 has over 50% repetitive content. We ran RepeatMasker (with species set to human, using the `rmblastn` algorithm) on the APG contigs, separately considering the placed and unplaced contigs. As might be expected since non-repetitive sequence is more easily anchored unambiguously, the placed contigs were not as repetitive overall as the unplaced contigs. RepeatMasker masks 61% of the placed sequence as repetitive, half of which (31%) was made up of simple repeats, with most of the remaining repetitive sequence made up of LINE, SINE, and LTR elements. The unplaced sequence was more repetitive, with an overall masking of 88%, where again the largest category, at 64%, was simple repeats, with another 22% classified as satellites (Figure 2.4). Mate information linking contigs to regions containing simple repeats or other repeat elements are less reliable links than those anchored in unique sequence, but still serve to provide tentative placements.

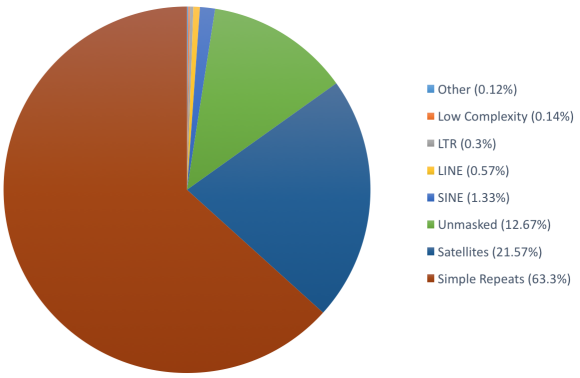


Figure 2.4 | African pan-genome repeat content. Breakdown of repeats in the APG contigs, placed and unplaced, as classified by RepeatMasker with species set to human and the rmblastn algorithm.

Additional screening and analyses

To screen for contaminants missed by Centrifuge, we used the Kraken metagenomics classifier[120] on our final set of representative contigs to compare them to a database containing all complete bacterial and archeal genomes, all viral genomes, selected fungi and protists, human, mouse, and known contaminant sequences. Any unclassified contig or contig hitting something other than mouse or human was further examined by running the blastn program[121] to align the contig to NCBI's non-redundant nucleotide database. We removed all contigs (as likely contaminants) that had alignments to a non-chordate covering greater than 50% of the contig with a BLAST e-value less than 10^{-10}. We additionally removed a single contig, also an apparent contaminant, hitting *Canis familiaris* at 90% identity over the entire contig, but lacking any strong matches to primates. As expected, all of these contaminant contigs were found in the set of unplaced contigs. We further examined the classifications of all removed contigs classified as viral or as *Plasmodium* to determine if any individuals appeared to have viral infections or malaria. All contigs with these classifications from Centrifuge or Kraken were screened for false positives by running blastn to align the contig to NCBI's non-redundant nucleotide database. Only contigs where all top BLAST hits covered at least 95% of the contig at an e-value less than 10^{-20} were considered to be true hits. This resulted in several contaminants of interest, including human betaherpesvirus and malaria. Since only assembled contigs were screened, all contaminants discovered were assembled into at least one contig of a minimum size of 1 kb with some individuals containing hundreds or thousands of assembled contaminant contigs, likely indicating a highly active infection. This resulted in the incidental discovery of 29 individuals with malaria infections and 1 with human betaherpesvirus (Table 2.2).

To ensure the final set of contigs were truly absent from the human reference genome, we re-aligned all APG contigs to GRCh38.p10 using bwa-mem[84] with default parameters. Two separate alignments were performed, one to the primary sequence and one to all patches and alternate loci. Although the reads used to assemble these contigs had initially failed to align to the genome, in some cases the resulting contigs had sufficient similarity that they could be aligned to primary sequence at or above 90% identity over at least 80% of the contig's length. This resulted in the removal of five two-ended placements, 24 one-ended placements, and 249 unplaced contigs. Among the 29 placed contigs that were removed at this step, all had alignments between 90% and 93% identity and were present in 10 or fewer samples; this may indicate that some individuals simply had slightly more divergent sequence than the overall population, explaining why Bowtie2 failed to align their reads initially. The best alignment locations that had at least 50% of the contig aligned to GRCh38.p10 at $\geq$ 80% identity were determined by taking alignments to both primary, alternate, and patch sequences, and calculating a score by multiplying the percent identity by the alignment length (Supplementary Tables 1-2 in Sherman *et al* 2019[78]). All placed locations were intersected with the NCBI provided gene annotations for GCF_000001405.36, which is the union of GenBank and RefSeq annotations for GRCh38.p10, and a translated BLAST search (blastx) was run against the NCBI protein database to identify potential protein-coding regions in the APG sequences.

Table 2.2 | **Contigs assembled from contaminants of interest.**

Sample ID	Population	*Plasmodium falciparum* (# contigs)	*Plasmodium malariae* (# contigs)	*Human betaherpesvirus 6B* (# contigs)
LP6005271-DNA_C04	Gabon	4184	-	-
LP6005271-DNA_D04	Gabon	3	-	-
LP6005271-DNA_E04	Gabon	2615	-	-
LP6005271-DNA_A03	Gabon	2	-	-
LP6005271-DNA_A04	Gabon	1	-	-
LP6005271-DNA_B03	Gabon	4077	-	-
LP6005271-DNA_C02	Gabon	6	2	-
LP6005271-DNA_C03	Gabon	2	-	-
LP6005271-DNA_F02	Gabon	36	-	-
LP6005092-DNA_B02	Nigeria	3	-	-
LP6005092-DNA_E03	Nigeria	8	-	-
LP6005092-DNA_H02	Nigeria	1	-	-
LP6005092-DNA_C02	Nigeria	1	-	-
LP6005092-DNA_A04	Nigeria	2	-	-
LP6005092-DNA_C01	Nigeria	1	-	-
LP6005092-DNA_G02	Nigeria	676	-	-
LP6005092-DNA_G03	Nigeria	89	-	-
LP6005092-DNA_H03	Nigeria	2	-	-
LP6005092-DNA_A06	Nigeria	1	-	-
LP6005092-DNA_B05	Nigeria	16	-	-
LP6005092-DNA_F06	Nigeria	15	-	-
LP6005092-DNA_A07	Nigeria	2	-	-
LP6005092-DNA_B06	Nigeria	7	-	-
LP6005092-DNA_B07	Nigeria	-	10	-
LP6005092-DNA_D06	Nigeria	5	-	-
LP6005092-DNA_E02	Nigeria	6	-	-
LP6005092-DNA_F05	Nigeria	1	-	-
LP6005092-DNA_G05	Nigeria	1	-	-
LP6005092-DNA_H06	Nigeria	1	-	-
LP6005107-DNA_F03	African American (Winston-Salem)	-	-	3

Contigs from *Plasmodium falciparum*, *Plasmodium malariae*, or *human betherpesvirus 6B* were assembled in 30 individuals. Though most *Plasmodium* contigs detected were falciparum, one individual had contigs present from both *Plasmodium* species and one solely from *malariae*. All individuals with *Plasmodium* contigs were from either the Gabon or Nigeria cohorts.

2.3 Genomic locations and analysis of 296 Mb non-reference sequence

In total, we discovered 296.5 Mb of novel DNA distributed across 125,715 sequences assembled from 910 individuals of African descent (Table 2.3, Figure 2.5). A total of 33,599 contigs with a combined length of 81,096,662 bases represented sequences present in at least two individuals in the CAAPA cohort. When alignments above 80% coverage and 90% identity to Chinese and Korean genome assemblies were also considered shared, the number of non-private insertions

Table 2.3 | Novel sequences in the African pan-genome.

	Number of sequence contigs	Total length (bp)	Bases with no alignment to GRCh38 (< 80% identity)	Longest contig (bp)
Two ends placed	302	667,668	431,656	20,732
One end placed	1,246	3,687,028	1,866,699	79,938
Unplaced	124,167	292,130,588	202,629,979	152,806
Total	125,715	296,485,284	204,928,334	152,806
Non-private only	33,599	80,098,092	50,044,650	152,806

Number and length of novel sequences in the African pan-genome. Bases with no alignment to GRCh38 were calculated by subtracting the lengths of all subsequences that aligned with at least 80% identity. The remainder represents truly novel sequence. Non-private insertions were insertions shared by at least two CAAPA cohort individuals.

increased to 61,410, totaling 160,475,353 bases and leaving 64,305 singleton contigs, a ~51% singleton rate. Of the 125,715 APG sequences, 1,548 (total length 4.4 Mb) were anchored to a specific location in the primary GRCh38 assembly. On average each individual contained 859 of these inserted sequences, with a single sequence being shared among 6 individuals (Table 2.4). Placed contigs were shared among more individuals, 196 on average, as shared sequences were more likely to meet the placement criterion in at least one individual.

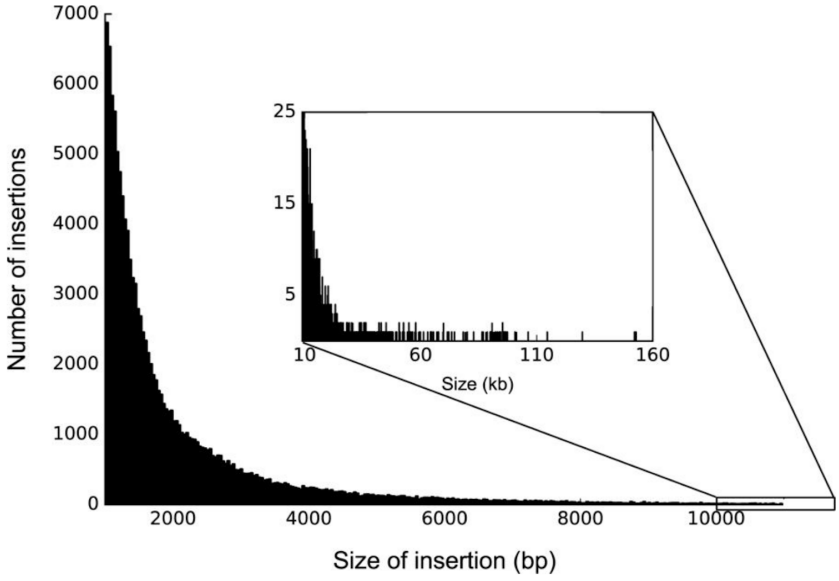


Figure 2.5 | APG contig length distribution. The 125,715 APG contigs range in length from just under 1 kb to 152.8 kb. All contigs under 1 kb are placed contigs which passed the 1 kb threshold pre trimming; the 219 contigs from 704 bp to 999 bp are not pictured.

We fully resolved the location for 302 of these sequences, and resolved the breakpoint of one end of the insertion for the remaining 1,246 (Supplementary Table 1 in Sherman *et al* 2019[78]). Placement locations were determined by complementing our methods with results from the PopIns program[102], which corroborated many placements and resolved placements for some insertions where our method was ambiguous. Of the 1,246 one-ended placements, our pipeline found 1,229, while PopIns [16] found an additional 17 and confirmed 60. Of the 302 pan-genome contigs for which both ends were placed, 70 were placed by both our method and PopIns, 129 by our method only, and 103 by PopIns only. Those placed solely by PopIns were verified via alignment of contig ends (as described above in Chapter 2.2).

Table 2.4 | African pan-genome contig presence/absence statistics.

	Number of contigs	Mean # insertions per individual	Mean # individuals per insertion
Two ends placed	302	120 (39.7%)	363 (of 910)
One end placed	1,246	212 (17.0%)	155 (of 910)
Unplaced	124,167	527 (0.4%)	4 (of 910)
Total	125,715	859 (0.7%)	6 (of 910)
Non-private only	33,599	758 (2.2%)	21 (of 910)

Statistics on the presence or absence of the African pan-genome contigs. Presence/absence was determined by aligning all raw contigs for each individual to the final set of APG contigs. Alignments of one or more contigs yielded a presence call if the alignments covered at least 80% of an APG contig at at least 90% identity. Additional presence calls were made for the placed contigs if the individual had a similar contig placed in the same location, even if the alignment thresholds were not met.

PopIns was able to resolve placement locations for some insertions where our method, which uses both contig and mate-pair alignments, gave ambiguous results. This is likely a result of PopIns' utilization of split-read alignments to determine exact placement locations, which provides some additional power. However, our approach has advantages when the contigs shared between individuals are more divergent, as they tend to be in the CAAPA populations. PopIns merges similar sequences prior to attempting placement but excludes from further analysis contigs which have a partial, but insufficient (for merging) match to many other contigs.

This resulted in PopIns considering only 5% of the full set of assembled contigs (81,650 of 1,536,049), while our approach placed many of these “unmergeable” contigs by placing all contigs first, and then clustering based on location. While this pre-merging step prior to placement worked well for the highly homogenous Icelandic population for which PopIns was designed, it was less effective for our more heterogenous African-descended populations.

The remaining 124,167 sequences could not be fully localized. However, mate-linking information pointed to a consistent location for at least one end for an additional 57,655 of these sequences (Supplementary Table 3). The longest placed sequence was 79,938 bp and appeared in 197 samples, and the longest unplaced sequence was 152,806 bp, which appeared in 11 samples (Table 2.3). Among all placed sequences, 387 intersected known genes, with placements within exons in 48 distinct genes, and within introns in an additional 267 genes. (Some genes contained more than one insertion.) Of the 315 genes containing insertions, 292 were named (i.e., had names other than "hypothetical" or a non-meaningful identifier). An additional 133 placed insertions and 46 already intersecting a protein coding gene intersected 142 distinct lncRNAs, 21 of which were named (Supplementary Table 4 in Sherman *et al* 2019[78]). A translated BLAST[121] search on unplaced sequences against NCBI’s nr database yielded an additional 10,667 contigs hitting a chordate protein with $\geq$ 70% identity and an e-value less than 10^{-10}. Placement locations and gene intersections were dispersed throughout the genome, and placed pan-genome elements were found on every chromosome (Figure 2.6), in addition to 115 insertions in chromosome-specific “random” sequence and 103 more in “unlocalized” sequences included in the primary assembly of GRCh38.

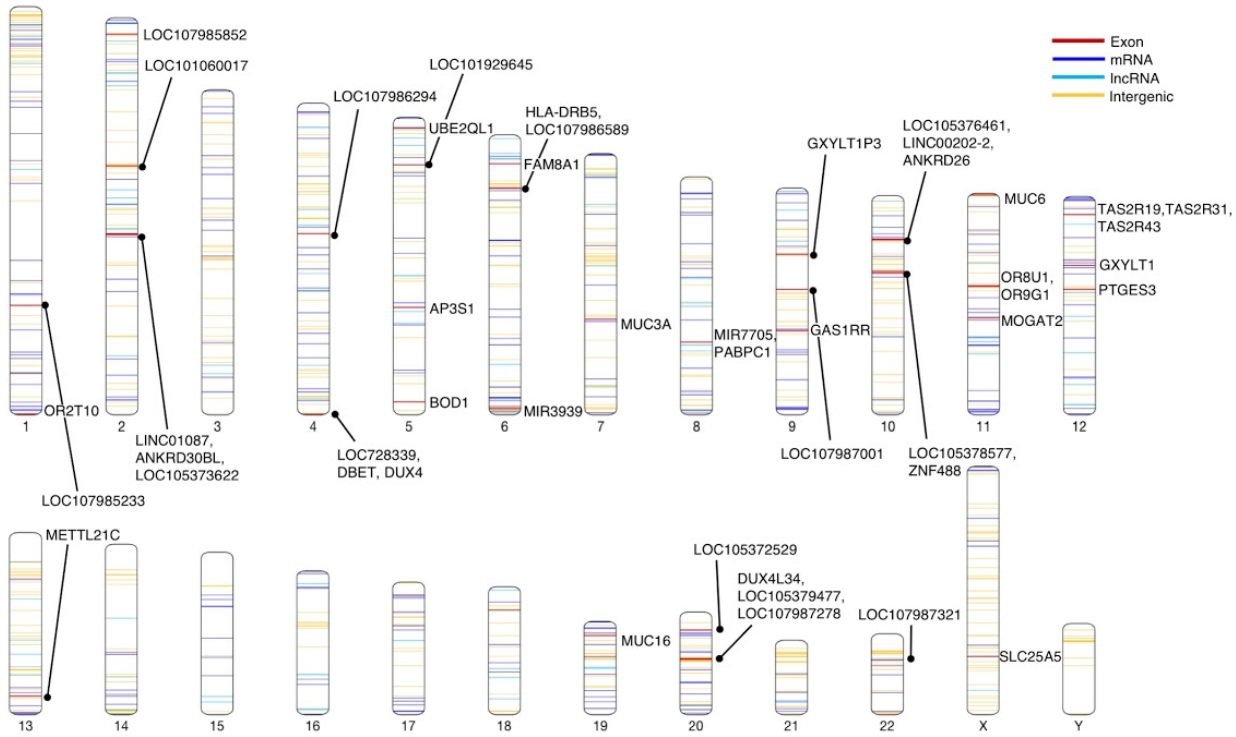


Figure 2.6 | African pan-genome contig locations. Map of the human genome showing the locations of all African pan-genome contigs, for those that could be placed accurately along one of the chromosomes. Yellow lines represent an intergenic location; blue lines represent insertion points with RNA but not exonic annotations, and red lines indicate intersections within exons. All exon-intersecting insertions are labeled with the gene name. mRNA and lncRNA gene names are reported in Supplementary Table 4 in Sherman *et al* 2019[78]. In some cases insertions are too close together for lines to be resolved; when this occurs within exons, gene names are listed in order by chromosome position. Line width is not to scale.

Of our APG contigs, 31,354,079 bases aligned to a GRCh38 “patch” or ALT locus as part of an alignment with an identity of ≥ 80%. An additional 60,202,871 bases aligned to the primary assembly at ≥ 80% identity; however, most of these alignments covered a small portion of an APG contig, and can be explained by the presence of extra copies of small repetitive elements. Supplementary Tables 1 and 2 in Sherman *et al* 2019[78] report alignments to ALT, patch, or primary assembly sequences covering at least 50% of the contig length with ≥ 80% identity. Requiring that at least 50% of a contig be aligned to any single location in GRCh38 produced a much smaller subset: of the 125,715 contigs, only 17,140 aligned to any part of GRCh38.p10

with a single alignment at ≥ 80% identity covering ≥ 50% of the contig length. These 17,140 contigs contain 22,420,979 aligned bases, with 13,770,950 bases being alignments to a reference chromosome. Although very few ALT loci in GRCh38.p10 are tagged with population-specific information, alignments of the CAAPA-specific sequences to these loci suggest an African source for some of these ALT sequences.

2.4 Presence of novel sequences in other genomes

We performed two different analyses to examine the presence of the APG sequences in other genomes. We performed comparisons to recent human assemblies of Chinese (HX1)[99] and Korean (KOREF1.0)[100] individuals, and several primate genomes by aligning our APG sequences to these assemblies using bwa-mem[84]. We additionally examined the presence or absence of the APG sequences in additional individuals sequenced with short-read WGS, from an independently sequenced cohort. To do this we used an identical alignment-assembly approach of assembling unaligned reads from these individuals, and compared the resultant contig sequences against the finalized APG contigs.

Comparisons to additional genome assemblies

We aligned all APG contigs to four additional genome assemblies: a Chinese genome assembly [14], a Korean genome assembly [15], the chimpanzee (*Pan troglodytes*) genome [30] (Genbank accession NC_006484, assembly GCA_000001515.7), and the rhesus macaque (*Macaca mulatta)* genome [31] (Genbank accession GCA_000772875.3). All alignments were performed

using bwa-mem with default parameters. Because bwa-mem sometimes found multiple distinct alignments for a contig, the best query-consistent set of alignments for each contig was retained, so no part of an APG contig aligned to more than one location in the reference. The best query-consistent set was determined by comparing the sums of alignment length weighted by percent identity. We then filtered these alignments to the Chinese, Korean, chimpanzee, and rhesus macaque genomes, retaining alignments with an overall identity ≥ 90% that covered ≥ 80% of the contig.

We compared each APG contig's alignment(s) on the Chinese and Korean genomes to all alignments of the same contig to GRCh38.p10 including patches and alternate loci, obtained as previously described. Among the contigs aligning to the Chinese or Korean genomes, we examined further those with a better alignment (higher identity x coverage) to the Chinese or Korean genome than to GRCh38.p10. We separated these further into two categories, those contigs with a "reasonably good" alignment to GRCh38.p10 (≥ 50% contig coverage and ≥ 80% identity for query-consistent sets of alignments within 1 kb of one another), and those lacking reasonably good alignments to GRCh38.p10. We compared alignments to the chimpanzee and rhesus macaque genomes with alignments to GRCh38.p10 in the same manner, as well as performing the same analysis comparing chimpanzee and rhesus macaque alignments to the Chinese and Korean alignments. For this analysis we used all query-consistent alignments to the Chinese and Korean genomes, not just those with ≥ 90% identity and ≥ 80% coverage.

We detected 42,207 contigs totaling 120.7 Mb aligning to either the Korean or Chinese assemblies with ≥ 90% identity and ≥ 80% contig coverage, and matching the Chinese or Korean assembly better than GRCh38. A vast majority of these contigs (32,955) had no alignment at ≥ 80% identity and ≥ 50% coverage to GRCh38.p10, indicating that these sequences were not simply divergent from GRCh38, but rather were not present at all (Table 2.5). An example of such a sequence and its alignments to GRCh38 and HX1 are shown in Figure 2.7a. This suggests these sequences have been lost in the small number of individuals used to create GRCh38, although some of them may reside in the few remaining gaps in the genome.

Table 2.5 | Comparison of African pan-genome contigs to the Chinese and Korean genomes.

	Best GRCh38 alignment is 80-90% identical with 50-80% coverage		Best GRCh38 alignment is < 80% identical or < 50% coverage		Total	
	Contigs	Length (bp)	Contigs	Length	Contigs	Length
Matches Chinese only	1,625	2,898,106	7,607	25,475,277	9,232	28,373,383
Matches Korean only	2,242	3,989,277	15,635	48,642,664	17,877	52,631,941
Matches both	5,385	9,720,662	9,713	29,981,048	15,098	39,701,710
Total	9,252	16,608,045	32,955	104,098,989	42,207	120,707,034

Contigs with a better alignment to the Chinese or Korean assemblies than to GRCh38. Alignments to the Chinese and Korean assemblies were required to have ≥ 90% identity and ≥ 80% coverage to be considered. Lengths shown are the sums of the contig lengths, not the alignment lengths.

While Shi *et al.* reported 12.8 Mb of novel DNA in the HX1 genome[99], we found a total of 68.1 Mb shared by HX1 and the unique sequences in the APG contigs (Table 2.5). To examine this further, we aligned the pan-genome sequences to the 12.8 Mb of novel sequence from HX1 and separately to the entire HX1 genome. We confirmed that the 68.1 Mb of sequences do align to HX1 (with ≥ 90% identity and ≥ 80% coverage for each contig), and these same sequences align poorly or not at all to GRCh38. For example, CAAPA_26854, a 15,617-bp APG sequence, aligned from positions 386–15617 (97.5% coverage) to HX1 Super-Scaffold_142 from positions

1338365–1353606 at 99.7% identity. To verify this sequence was novel, we aligned it to GRCh38.p10 using BLAST [26] and nucmer [34] in addition to bwa-mem [28]. The best match was an alignment of just 425 bp at 81.9% identity, demonstrating that this sequence is essentially absent from GRCh38. However, Super-Scaffold_142:1338365-1353606 was not included in the 12.8 Mb of novel sequence reported as unique to the Chinese genome [14]. This reporting discrepancy is methodological: the Chinese genome assembly has relatively large scaffolds which were considered unique only if a large proportion of the scaffold failed to align to GRCh38.

Additional comparisons to the chimpanzee and rhesus macaque genomes were performed to help ensure the APG sequences were not contaminants. Because ~90% of the human genome can be aligned to chimpanzee at > 98% identity, we expected much of the pan-genome should be detectably similar to chimpanzee. After filtering to retain only query-consistent alignments, 123,582 contigs had some portion aligning to chimpanzee, with the alignment lengths totaling 177,388,896 bases. Although the alignment lengths to chimpanzee only make up ~60% of the total APG sequence, an additional 17,544,523 bases that did not align to chimpanzee aligned to rhesus macaque, for a total of 194.9 Mb on 123,603 contigs aligning to at least one non-human primate. Furthermore, over 98% of APG contigs had at least a partial alignment to chimpanzee or rhesus macaque, providing further validation that these sequences are human in origin rather than contaminants. Additionally, near-perfect alignments to chimpanzee intersecting known genes may provide additional clues about the functionality of these sequences. For example, the novel sequence in one of our placed contigs contained an exon of the KDMB6

gene in chimpanzee, though this exon is missing from the GRCh38 annotation (Figure 2.7b). It appears that this exon is present in some humans as well as in chimpanzee. This KDMB6 insertion has since been validated by the discovery of this variant in independent datasets including TOPMed[68] and Audano *et al*[27], both of which report finding this insertion in all of their sequenced individuals, indicating that the reference sequence is either a rare deletion or an error, and this insertion is in fact the common, and ancestral, human variant.

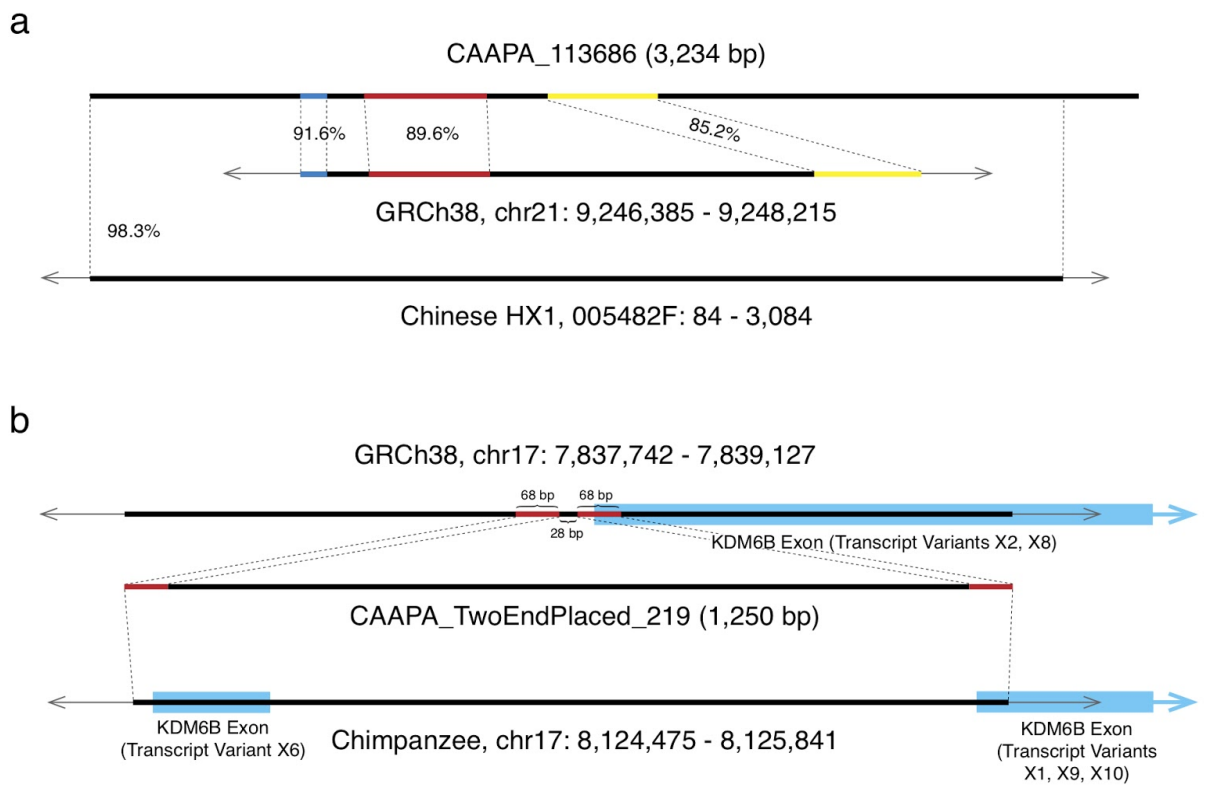


Figure 2.7 | APG alignments to other genomes. (a) An example of an alignment which does not meet the 50% coverage, 80% identity threshold for a "reasonably good" alignment to GRCh38. The APG contig is shown at the top, with the best consistent alignments to GRCh38 in the middle. The three constituent alignments (blue, red, and yellow segments) cover 801 bases, just under 25% of the contig, with a cumulative weighted identity of 87.9%. CAAPA_113686 has a single near perfect alignment to a Chinese HX1 contig (delineated by dotted lines) covering over 80% of CAAPA_113686 at over 90% identity. The APG contig also aligns very well to Korean and chimpanzee contigs (not shown). **(b)** Detailed alignments of a 1250-bp APG contig (1386 bp untrimmed) to GRCh38 and chimpanzee (Pan_tro 3.0). The alignment of both ends indicates the contig's position on human chromosome 17 (top), where it is missing from GRCh38. Red segments represent the aligned regions, which were trimmed prior to reporting the final APG sequence. CAAPA_TwoEndPlaced_219 aligns end-to-end to chr17 of the chimpanzee genome (bottom), where it intersects two annotated exons of the KDM6B gene. On GRCh38 the exon beginning after the inserted sequence is the first exon of the gene.

Of the 123,582 contigs with some alignment to chimpanzee, just 668 contigs (1,666,093 bases, or 0.56% of the total APG sequence) aligned better to chimpanzee than to the three human assemblies (and aligned to chimpanzee with ≥ 90% identity and ≥ 80% coverage), and 329 of these (934,222 bp) were entirely missing from the three human assemblies (Table 2.6). Comparisons to rhesus macaque yielded an additional 18 contigs (50,174 bp) aligning better than to all three human assemblies, 4 of which (9,403 bp) were missing from the three human assemblies. These sequences, while present in multiple individuals from the CAAPA collection, may have been lost in other lineages. It is also possible the sequences are present in those genomes but simply missing from the assemblies, all of which contain many gaps.

Table 2.6 | African pan-genome contigs with better alignments to chimpanzee than human.

Pan-genome contig aligns better to chimpanzee than:	Human alignment ≥ 80% identity and ≥ 50% coverage		Human alignment < 80% identity or < 50% coverage		Total	
	Contigs	Length (bp)	Contigs	Length	Contigs	Length
GRCh38	1,184	2,207,895	1,466	3,807,254	2,650	6,015,149
Chinese genome	476	1,074,143	480	1,298,145	956	2,372,288
Korean genome	633	1,373,570	448	1,266,471	1,081	2,640,041
All 3 human assemblies	--	--	329	934,222	668	1,666,093

African pan-genome contigs that align better to chimpanzee than to the GRCh38, Chinese, or Korean genomes. All contigs had alignments to chimpanzee with ≥ 90% identity and ≥ 80% coverage. Lengths shown are the sums of the contig lengths.

As an additional check to ensure the APG sequences were not contaminants, we examined what portion of contigs had some match, even just a partial one, to the GRCh38, Korean, or Chinese assemblies. After filtering to retain only query-consistent alignments, 98% of the contigs (123,600) had some portion aligning to either the Chinese, Korean, or GRCh38 assemblies. The Korean assembly had the most alignment, with 123,585 contigs contained an alignment totaling

247.2 Mb of aligned length, or 83% of the total APG sequence, although only 31,033 contigs, totaling 80.9 Mb of alignment, aligned with over ≥ 90% identity and ≥ 80% coverage.

Determining presence/absence of APG contigs in WGS-sequenced samples

We additionally called presence/absence of the APG insertions in short-read sequencing data from 12 individuals from 6 European populations and 12 individuals from 6 continental African populations from the Simons Genome Diversity Project (SGDP) [27]. As with the CAAPA individuals, short reads from these 24 individuals were aligned to the reference genome, and unaligned reads assembled with MaSuRCA. Raw contigs from the MaSuRCA assemblies (including contigs under 1 kb) were aligned to the final set of APG contigs with bwa mem using default parameters. Alignments to an APG contig aligning within 300 bp of one another were chained to create longer alignments where possible. Identity of the chained alignment was taken to be the identity of these alignments weighted by length, and coverage was taken to be the total aligned bases over the total APG contig length. If an individual's raw contig alignments produced an alignment with ≥ 90% identity and ≥ 80% coverage to an APG contig, that APG contig was called as present.

The SGDP samples varied widely in the number of APG sequences they contained; 4 of the Africans and 4 of the Europeans contained ~1000 APG sequences each, while 5 Africans and 1 European (English) sample contained ~700 insertions (Figure 2.8). This could be due to admixture in the CAAPA samples, in which 9 of 19 cohorts were African-American, or to admixture in the European SGDP samples, or some combination of both.

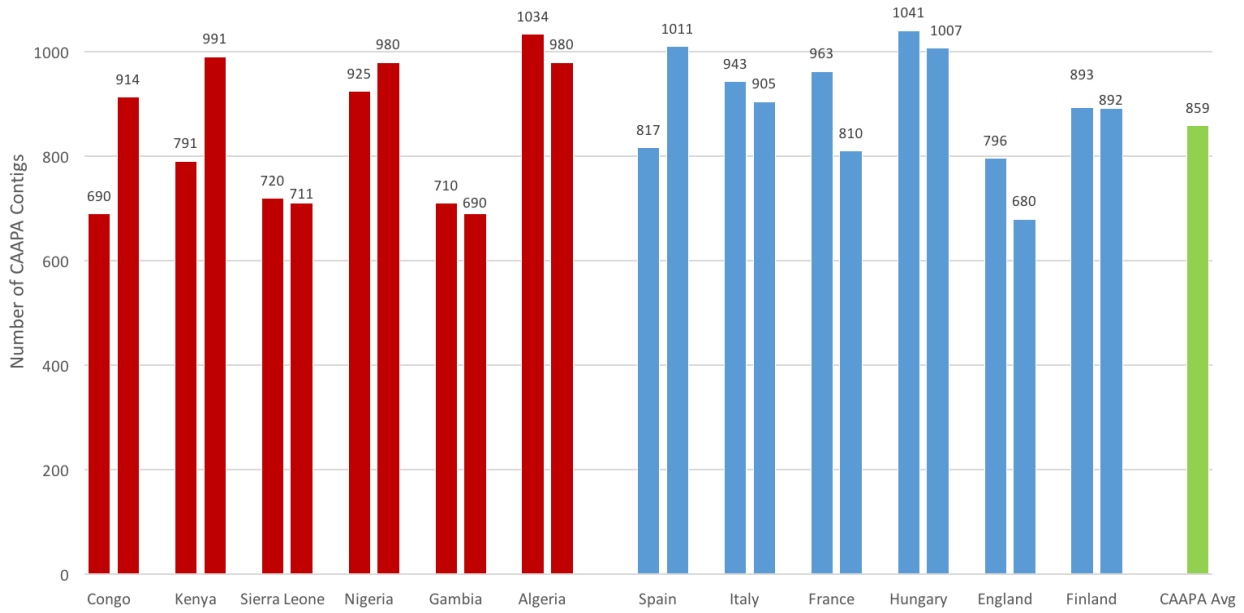


Figure 2.8 | APG contigs present in Simons Genome Diversity Project populations. Twenty-four individuals from the Simons Genome Diversity Project from 12 populations, 6 African (red) and 6 European (blue), were examined to determine presence/absence of the APG contigs. Each individual's assembled contigs were aligned to the APG contigs to determine the number of APG contigs present in the individual. The same analysis was performed to determine, via this genotyping method, the average number of contigs per CAAPA cohort individual (green).

To further examine how well the APG contigs represent continental African populations, we additionally examined the APG contigs present in only the continental African SGDP samples and only the European SGDP samples, but not both. The European SGDP samples cumulatively contained 4,645 of the APG insertions, while the African samples contained 4,381 insertions, with 1,961 of these insertions appearing in both populations. We examined the 2,684 present in the 12 European samples but not the African samples, as well as 2,420 present in the African but not European samples. Although these sequences may not be fully specific to these populations because we examined only 12 samples, we took these to represent sequences tending toward European and African specificity. The European-specific contigs were found at a somewhat lower frequency in the CAAPA samples (47 contigs per individual on average) than

the African specific contigs (63 per individual on average). This difference provides some evidence that despite the admixed nature of the dataset, the APG sequences reported represent sequences present in African populations more so than European populations, though the inclusion of European derived DNA is expected in the APG sequences due to admixture.

The same alignment procedure was performed for all 910 CAAPA individuals as well to determine presence/absence of each APG contig in each individual. This genotyping was done after finalizing the APG insertion set both due to the possibility that an APG insertion originally assembled in multiple pieces under 1kb, and thus was not initially included, but was indeed present in a CAAPA individual, as well as due to the difficulty of tracking best alignments during the stage where contigs from all 910 individuals were iteratively merged. We report this presence/absence genotyping in the CAAPA individuals as a matrix of contigs by individuals, where a "1" was included in the matrix if the alignment criteria described above were met (Supplementary Data in Sherman *et al* 2019[78]).

Additionally, for the placed contigs, because we had already determined which individuals contained these sequences, the genotype matrix was supplemented by adding a presence call ("1") if we had determined that an individual had a contig in the placement cluster. This additional calling allowed for increased sensitivity for individuals who had mate placement information available for the insertion, even when the contigs did not meet the identity/coverage criteria used for genotyping. The "genotype" matrix entries indicate

presence/absence calls represented as 1 or 0; heterozygous and homozygous genotypes are not differentiated.

Using the genotype matrix, we estimated whether the pan-genome would continue to grow as more individuals were sequenced by randomly sampled varying numbers of individuals within our dataset and using the genotype matrix to determine, in each subset, how much of the APG sequence was present. Each data point was an average of 10 random samplings, each with the same number of individuals. The amount of DNA added to the pan-genome appears to be increasing approximately linearly as the sample size grows, and has not reached an asymptote with 910 individuals, though the amount of shared sequence begins to level off, indicating that as more individuals are sequenced, the number of private insertions grows more dramatically than shared sequences (Figure 2.9).

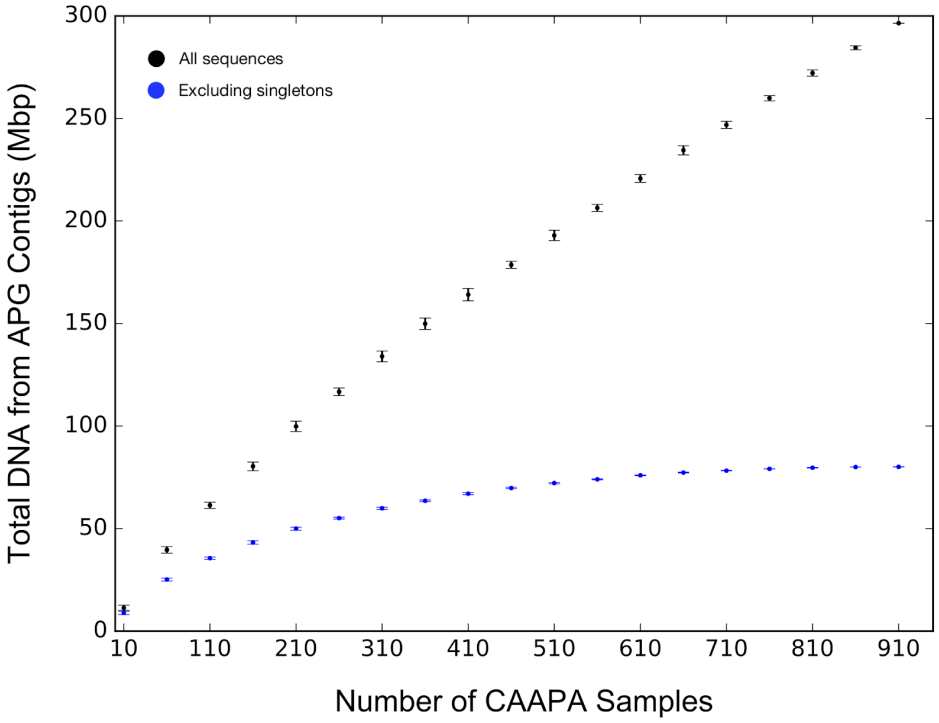


Figure 2.9 | Amount of APG DNA by number of individuals. Individuals were randomly sampled and the combined lengths of the set of APG contigs contained in those individuals was recorded, as determined by the presence/absence genotyping matrix. Each point is an average of 10 random samples of the given number of individuals. Error bars are standard deviation. For the blue markers, only DNA present in more than one of the 910 individuals is counted in the total.

2.5 Implications of discovering 296 Mb of non-reference sequence in 910 individuals

Our findings here demonstrate that the standard human reference genome lacks a substantial amount of DNA sequence compared to other human populations. The APG sequences contain 296.5 Mb, equal to 10% of the genome, regions that will necessarily be missed by any efforts relying only on GRCh38 to study human variation, as nearly all studies do today. Of these 296.5 Mb, 120.7 Mb were shared by the Korean or Chinese populations, suggesting those regions may have been lost more recently or may be rare in the specific populations represented in GRCh38. Additional analyses of the APG sequences have also examined satellite repeats. While we reported via RepeatMasker that satellite DNA comprises ~22% of the sequences, a subsequent, more focused analysis by Karen Miga determined that ~85% of the APG sequences have exact oligo matches to HSAT II or III repeats[73]. The variation in HSAT II and III repeats between individuals and populations is understudied, in part because of a lack of representation of these sequences in GRCh38, though these centromeric satellites have been implicated in disease association studies. The APG sequences we present here can aid in examining HSAT variation, and be a first step to producing better references for these diverse regions.

Overall these results suggest that a single reference genome is not adequate for population-based studies of human genetics. And although we are amassing a wealth of pan-genomic data in both global and population-specific studies such as ours, what to do with these data remains an open question. The creation of a single, global human pan-genome holds conceptual appeal, and cataloguing all human variation is a noble goal. However, to date no computational method is capable of aligning human sequences to a pan-genome of all human

variation, while enforcing that alignments be biologically plausible, although research on efficient indexing, storage, and traversal of graphical representations is actively underway that might solve this problem[122–124].

Instead, a better approach may be to create reference genomes for all distinct human populations, which over time will eventually yield a comprehensive pan-genome capturing all of the DNA present in humans. Efforts to develop additional reference genomes are gaining steam in recent years, in part because our ability to accurately assemble these genomes is rapidly improving. Recently the first telomere-to-telomere assembly of a human chromosome[125] demonstrated that newly produced genomes utilizing third-generation sequencing have the potential to be of higher contiguity than the current reference. We recently produced an assembly of an Ashkenazi individual with contiguity surpassing GRCh38 and complete gene annotation[101], using PacBio and Nanopore data and a recent paper produced an assembly of a widely used cell line, WI-38 using PacBio data, which contains 153 scaffolds over 1 kb which align to our APG contigs[126], demonstrating that better assemblies will capture some of the novel sequences we have reported which are present in many individuals but absent from GRCh38. With a plethora of human reference genomes, individuals could be analyzed by comparing them to their closest matching population or populations, even if this information is not known a priori, resulting in far fewer 'missed' sequences when aligning an individual to a reference genome.

Regardless of what representations are ultimately used to capture these genomes, it seems inevitable that we will soon move beyond our reliance on a single human reference genome, which as we have demonstrated in our African pan-genome analyses, is not sufficiently representative of human populations. Approaches that capture the vast amounts of variation in the population, whatever computational form they may end up taking, will be a critical tool in helping us understand and analyze the genetic instructions that make us human.

2.6 Commands and parameters

Bowtie 2 alignment, per sample

```
bowtie2-build [GRCh38_no_alt] [GRCh38_no_alt_idx]
bowtie2 -x [GRCh38_no_alt_idx] [reads1] [reads2] > [alignments.bam]
```

Extraction of unaligned reads (and mates) via samtools, per sample
---

```
samtools fastq –f 12 [alignments.bam] -1 [mateUnmapped_R1.fq] -2
[mateUnmapped_R2.fq]
samtools fastq –f 68 –F 8 [alignments.bam] > [R1_mateMapped.fq]
samtools fastq –f 132 –F 8 [alignments.bam] > [R2_mateMapped.fq]
samtools view –f 8 –F 4 [alignments.bam] > [GRCh38Links.bam]
```

MaSuRCA assembly, per sample

```
masurca_config.txt:
#########################################################
DATA
PE= pe 300 50 [mateUnmapped_R1.fq] [mateUnmapped_R2.fq]
PE= s1 300 50 [R1_mateMapped.fq]
PE= s2 300 50 [R2_mateMapped.fq]
END

PARAMETERS
GRAPH_KMER_SIZE=auto
USE_LINKING_MATES=1
KMER_COUNT_THRESHOLD = 1
NUM_THREADS=24
JF_SIZE=200000000
```

```
DO_HOMOPOLYMER_TRIM=0
END
************************************************************

masurca masurca_config.txt && ./assemble.sh
```

Centrifuge, per sample

```
centrifuge --report-file [centrifuge.report] -x [centrifugedb] -k 1
--host-taxids 9606 -f [masurca_contigs_over1kb.fa] > [centrifuge.output]

centrifuge-kreport -x [centrifugedb] [centrifuge.output] --min-score 0
--min-length 0  > [centrifuge.krakenOut]
```

*** `centrifuge.krakenOut` was used to filter any non-chordate identified reads. ***

RepeatMasker on assembly contigs, per sample

```
RepeatMasker -nolow -species human [filteredContigs.fa]
```

Bowtie 2 alignment of reads to contigs, per sample

```
bowtie2-build [filteredContigs.fa.masked] [contigIdx]
bowtie2 -x [contigIdx] -U [R1_mateMapped.fq],[R2_mateMapped.fq] -S
[readContigAlignment.sam]
```

Linking mates to implicated region, and aligning to region, per sample

---

```
samtools view -h -F 256 [readContigAlignment.sam] | samtools sort - -n –O bam
| bedtools bamtobed -i stdin | awk '{OFS="\t"}{print $4,$1,$6,$2,$3}' | sort >
[readContigAlignment.txt]

samtools view -H [GRCh38Links.bam] | cat - <(awk 'FNR==NR{main[$1]=$0;next} $1
in main {print main[$1]}' <(samtools view [GRCh38Links.bam])
[readContigAlignment.txt]) | samtools sort -n –O bam | bedtools bamtobed -i
stdin | awk '{OFS="\t"}{print $4,$1,$6,$2,$3}' | sed -e 's/\/[1-2]//g' | sort
> [matchedMates.txt]

join -j 1 [readContigAlignment.txt] [matchedMates.txt] > [mateLinks.txt]
```

*** Filtering was performed here using python scripts to examine links to contig ends only, and filter based on described unambiguity criteria as described in Chapter 2.2. Contig ends and GRCh38 regions meeting criteria were extracted with `samtools faidx` ***

```
nucmer --maxmatch -l 15 -b 1 -c 15 -p [deltaFile] [GRCh38Regions.fa]
[filteredContigEnds.fa]
```

Clustering of placed contigs

```
bedtools merge -d 100 -c 4 -o distinct [placedCtgLocations.bed] >
[mergedClusters.bed]
nucmer -p [deltaFile] [repCtg.fa] [restOfClusterCtgs.fa]
nucmer -p [deltaFile] [verifiedClusterCtgs.fa] [unplacedCtgs.fa]
```

Left/Right one end placement merging into two end placement

```
nucmer --maxmatch --nosimplify -p [deltaFile] [leftEndedPlaced.fa]
[rightEndPlaced.fa]
show-coords -H -T -l -c -o [deltaFile] > [coordsFile]
```

Removal of redundant placements

```
nucmer --maxmatch --nosimplify -p [deltaFile] [allPlaced.fa] [allPlaced.fa]
```

Clustering of unplaced contigs

```
nucmer --maxmatch --nosimplify -l 31 -c 100 -p [deltaFile] [unPlaced.fa]
[unPlaced.fa]
show-coords -H -T -l -c -o [deltaFile] > [coordsFile]
```

*** Additional analysis was performed on the alignments to find and remove contigs contained within two contigs with the ends overlapping (see Chapter 2.2) ***

Further screening

```
kraken --db [database] [APG_Sequences.fa]

blastn -db [nt] -query [kraken_nonMammalHits.fa] -outfmt "6 qseqid sseqid
pident length mismatch gapopen qstart qend sstart send qlen slen evalue
bitscore qcovs qcovhsp staxids sscinames" -max_hsps 1 -max_target_seqs 1 -out
[blastOutput]

bwa index [GRCh38.p10_primaryChrs]
bwa index [GRCh38.p10]
bwa mem [GRCh38.p10_primaryChrs] [APG_Sequences_noContamiants.fa]
bwa mem [GRCh38.p10] [APG_Sequences_noContamiants.fa]
```

Genotyping, per sample

```
bwa index [APG_Sequences_final.fa]
bwa mem [APG_Sequences_final.fa]  [contigsFromMaSuRCA.fa] >
[sampleToAPGAlignment.sam]
```

Comparisons to other genomes

```
bwa index reference
bwa mem [reference] [APG_Sequences_final.fa]
```

Reference genomes used

```
GRCh38_no_alt    GCA_000001405.15_GRCh38_no_alt_analysis_set.fna
   GRCh38.p10    GCF_000001405.36_GRCh38.p10_genomic.fna
        KOREF    GCA_001712695.1_KOREF1.0_genomic.fna
          HX1    hx1f4s4full_3rdfixedv2.fa
   Chimpanzee    GCA_000001515.7
       Rhesus    GCA_000772875.3
      Macaque
```

In addition to these primary commands, additional filtering steps and custom analyses were performed, as described previously. Filtering commands were primarily performed using `awk` and filtering for identity and coverage was always performed on `coords` files produced by `show-coords`; if `bwa` was used for alignments in place of `nucmer`, the `sam` files were converted to `nucmer` delta files, using the CIGAR strings and lengths to determine identity and coverage.

2.7 Addendum

A recent BioRxiv paper performed additional analyses on the African pan-genome sequences described here to determine if they contain any non-human contaminants[127]. This analysis went beyond our sequence-based filtering, using a combination of translated nucleotide searches against a protein database with Kaiju[128], plus additional DIAMOND[129], BLASTn[121], and protein prediction/family identification on the non-repetitive sequences in the APG contigs. The paper reports 1,475 APG contigs as contaminants. 933 of these were additionally previously reported as contamination in a large-scale analysis of GenBank and RefSeq, where over 2 million contaminated GenBank entries were found, including 1,003 contigs from the APG set[130]. We have since updated the GenBank entry for the APG sequences, removing all 1,475 contigs reported as contamination in the Manni and Zbobnov BioRxiv paper. The results in Chapter 2 have not been updated since this removal, so the numbers presented are slightly different than they would be if recalculated with the updated APG set. However, as only ~1% of the total APG sequence was removed, the overall findings and conclusions from Chapter 2 remain unchanged.

Chapter 3: Utilizing RNA-seq to discover novel exons in non-reference sequences

Some of the work in Chapter 3 was performed by my summer undergraduate mentee, William Cho, under my supervision.

3.1 Background: RNA-sequencing and analyses

RNA can be sequenced in a similar manner to DNA; mRNA is taken from the cell and used as a template strand to create cDNA (complementary DNA), which can be sequenced as described in Chapter 1.2. Recent developments also allow for sequencing RNA from a single cell, and direct sequencing of RNA, but these advances will not be discussed here. The majority of RNA sequencing (RNA-seq) is RNA taken from many cells at once (bulk RNA-seq), though tissues are generally separated, and RNA-seq is often used to examine tissue specific expression. Since in humans, genes contain introns and exons, and the introns are spliced out to create mRNA, aligning sequenced mRNA back to the human genome will necessarily not align in one piece; RNA-seq will only align to exonic regions, and reads spanning the splicing boundary between exons will align non continuously in a split manner, half to the end of one exon, half to the start of the next, with an intron between. To account for this, specialized spliced aligners are used for RNA-seq data, including STAR[131], TopHat2[132], and HISAT2[63], which consider that these alignments split across exons should be 'good' alignments. These aligners can take in gene annotation information on the reference genome to provide an additional prior for where read splits are expected, though the aligners can find novel splits as well, where exons are not annotated.

An additional set of tools, transcriptome assemblers, such as Scallop[133], or StringTie2[134], can take in spliced RNA-seq alignments, and attempt to resolve what isoforms of a transcript are present. In humans, a phenomenon called alternative splicing allows a single gene to be spliced in various ways; including and excluding different exons in a given transcript. This is thought to allow for more flexibility -- many more isoforms, and thus proteins, are possible than just the number of genes. Short read lengths, however, make transcript assembly difficult. Exons are linked together based on the presence of reads spanning the boundary of two exons, but seeing for example, reads spanning exon 1 and 3, and reads spanning exon 3 and 4, does not necessarily mean that a transcript with exons 1, 3, and 4 exists -- the spanning reads may belong to different transcripts. Reads spanning more than two exons are rare with short read sequencing, as exons are typically longer than the read length. Transcriptome assembly can additionally consider coverage information to determine what transcripts exist; if transcripts are expressed at differing levels, the number of reads spanning each exon junction can be examined to determine the most statistically likely pattern of exons/transcripts.

RNA-sequencing data is abundant. Data sets exist across many populations, body tissues, and across changing conditions and time points within a single individual's life. One of the most extensive RNA-seq datasets is from the Gene-Tissue Expression (GTEx) Consortium, which has produced RNA-seq data from over 50 tissues due to post-mortem collection, as internal tissue samples are difficult to impossible to obtain from live donors, and has over 17,000 samples in the latest version[135]. Much like with whole genome sequencing data, the typical pipelines for analysis, as described above and illustrated in Figure 3.1, begin with alignment to the reference

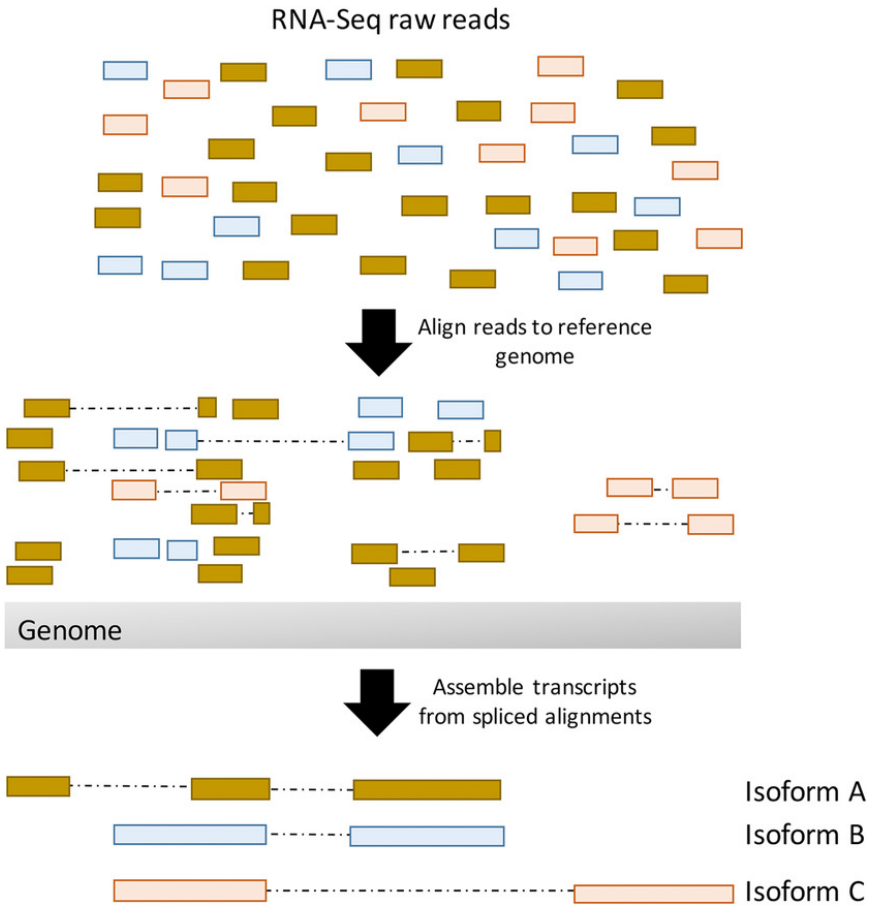


Figure 3.1 | A typical RNA-seq analysis pipeline. First, RNA-seq reads are aligned to the reference genome with a spliced aligner. Spliced alignments are indicated by dashed lines. Then, a transcriptome assembler attempts to resolve the alignments into isoforms. Here, there are three isoforms, the 'truth' of which isoform a read originated from is indicated by color, but an assembler may not be able to resolve this, as there are no spliced links between, for example, the two exons with pink reads aligned to them. Figure from Costa *et al* 2018[136].

genome. Thus as with reference-based DNA-sequencing analyses, this approach will miss any novel transcribed sequences. Furthermore, RNA-seq analysis is highly biased by the existing gene annotation, which is often provided to both alignment and transcript assembly tools (and for some tools is required). While providing annotation can help with downstream expression quantification on known genes and transcripts, it will at best bias tools away from finding new genes and transcripts, and at worst will be unable to find them at all, in the case of tools which only perform reference-guided transcriptome assembly. Recent research has shown that despite many years of examining the human genome through RNA-seq data, we still continue to find

new transcripts present in reference sequence and 'known' transcripts that appear to be simply noise[71]; what transcription we might find in non-reference sequence is still largely unexplored.

3.2 Analysis of 296 Mb non-reference sequence for transcription potential

While the location of much of the 296 Mb of non-reference sequence, described in Chapter 2, is unknown, it is possible that some of the novel sequence is transcribed. Unfortunately, we don't have RNA-seq data from the individuals these sequences came from, but RNA-seq data is available from many thousands of other individuals through the GTEx consortium. Since the novel sequences are expected to be present in many individuals based on our analyses of sequence present in data from the Simons Genome Diversity Project[52] and the Korean[100] and Chinese[99] reference genomes, we can look at public RNA-seq data for and see if any individuals and tissues appear to have these sequences transcribed. To do this we can align GTEx RNA-seq data, which in the latest version, v8, consists of over 17,000 samples from 948 post-mortem donors across 54 tissues[135], to the novel sequences. Indications of transcription in the African Pan Genome (APG) sequences would include:

1. Spliced alignments of RNA-seq reads to APG contigs with
    a. Deep coverage of these spliced alignments
    b. Spliced alignments in multiple individuals and multiple tissues
    c. Clear differential expression across tissues
2. Alignments are predominantly not from multi-mapped reads
3. Regions the reads align to are not repetitive/low-complexity
4. Regions with alignments have matches to known mammalian proteins when translated

5. Regions with alignments align to annotated genes in primates or other mammals

We have performed a preliminary analysis, using a subset of 93 GTEx samples, 3 samples in each of 31 tissues, to examine these questions and decide whether to proceed further using the full GTEx data set. Reads from each of the 93 samples which did not align to GRCh38 were aligned to the 125,715 APG sequences, using TopHat2[132] (a HISAT2[63] bug at the time, which has since been resolved, led us to use TopHat2). StringTie2[134] was then run in a per contig (pooling reads from samples) manner, as in many cases a single sample did not have sufficient coverage for StringTie2 to produce a novel transcript, even where splice alignments existed. Contigs were then analyzed and using a python script a number of features were reported per contig. These features included the number tissues with alignment, the total number of reads aligning and the percentage of those aligning in a spliced manner, information from StringTie2 (number of transcripts, number of exons), how many of the potential exon regions were masked with RepeatMasker[117], and a breakdown of read mapping quality. This information was provided for pooled sample data per contig, as well as reported for the single sample with the most aligned reads to the contig. By using this method, different metrics can easily be used to determine the 'best' candidates, which can then be examined by hand, prior to proceeding and determining how to prioritize candidates in a more principled manner. A sample of the csv output produced by this pipeline, sorted by the percentage of spliced reads, is shown in Figure 3.2.

Contig	Total # Tissues	Tissues	Total Reads	Spl Reads	% reads spliced	Non-Spl Reads	MapQ	Contains Masked	Human Annotation(s)	# Transcripts	# Exons	Exon Summary
CAAPA_OneEndPlaced_1042	8	['Salivary_Gland : 0.9104	949	902	0.950	47	['3.0 : 0.376185	FALSE	['mRNA : MUC19'	2	17	721:771__910:963__194
CAAPA_OneEndPlaced_337	10	['Salivary_Gland : 0.9103	803	758	0.944	45	['3.0 : 0.398505	FALSE	['mRNA : MUC19'	1	7	2139:2196__2317:2370_
CAAPA_OneEndPlaced_847	5	['Salivary_Gland : 0.8593	128	118	0.922	10	['1.0 : 0.507812	FALSE	['mRNA : MUC19'	1	3	839:998__1137:1190__2
CAAPA_107648	20	['Salivary_Gland : 0.2889	3703	1791	0.484	1912	['0.0 : 0.874426	FALSE		2	8	5150:5272__6143:6267_
CAAPA_TwoEndPlaced_218	23	['Salivary_Gland : 0.2815	5474	2619	0.478	2855	['0.0 : 0.843624	FALSE	['mRNA : MUC16'	4	18	1881:2103__2224:2456_
CAAPA_OneEndPlaced_1051	19	['Salivary_Gland : 0.2851	2195	1031	0.470	1164	['0.0 : 0.984054	FALSE	['mRNA : MUC16'	1	3	6604:6639__6900:6965_
CAAPA_125514	19	['Salivary_Gland : 0.3013	2270	1056	0.465	1214	['0.0 : 0.973127	FALSE		1	3	5742:5868__7165:7230_
CAAPA_114586	23	['Salivary_Gland : 0.2971	4243	1922	0.453	2321	['0.0 : 0.794249	FALSE		5	17	1877:2093__2411:2446_
CAAPA_51080	19	['Salivary_Gland : 0.3123	2523	1119	0.444	1404	['0.0 : 0.947284	FALSE		1	3	5751:5877__7177:7242_
CAAPA_5697	25	['Salivary_Gland : 0.2822	4163	1842	0.442	2321	['0.0 : 0.803747	FALSE		2	11	5517:5582__6452:6576_
CAAPA_83523	18	['Salivary_Gland : 0.3032	2361	1039	0.440	1322	['0.0 : 0.977128	FALSE		2	5	1137:1209__2080:2251_
CAAPA_65303	19	['Cervix_Uteri : 0.358078	1145	470	0.410	675	['0.0 : 0.870742	FALSE		2	6	637:725__1596:1720__1
CAAPA_62443	9	['Salivary_Gland : 0.2941	306	124	0.405	182	['0.0 : 1.0']	FALSE				
CAAPA_19880	20	['Salivary_Gland : 0.2985	2857	1150	0.403	1707	['0.0 : 0.953797	FALSE		1	3	6673:6799__8095:8160_
CAAPA_94747	16	['Salivary_Gland : 0.3386	629	243	0.386	386	['0.0 : 0.976152	FALSE				
CAAPA_110953	19	['Salivary_Gland : 0.2874	1343	435	0.324	908	['0.0 : 0.973194	FALSE				
CAAPA_29100	15	['Salivary_Gland : 0.3142	1365	389	0.285	976	['0.0 : 0.944322	FALSE		1	3	2485:2520__2781:2846_
CAAPA_114551	15	['Salivary_Gland : 0.2948	485	110	0.227	375	['0.0 : 0.948453	FALSE		1	2	1087:1122__1447:1619
CAAPA_60818	10	['Salivary_Gland : 0.3731	544	111	0.204	433	['0.0 : 0.976102	FALSE		1	2	1320:1492__1826:1854
CAAPA_35603	12	['Salivary_Gland : 0.2828	449	76	0.169	373	['0.0 : 0.995545	FALSE				
CAAPA_81614	28	['Testis : 0.980601282943	25878	2075	0.080	23803	['50.0 : 0.90903	TRUE		2	4	1:480__1:733__774:1064
CAAPA_78986	30	['Spleen : 0.29320206282	10665	372	0.035	10293	['0.0 : 0.718799	TRUE		2	3	1:558__26:286__586:105
CAAPA_80794	30	['Spleen : 0.28109517601	7670	249	0.032	7421	['0.0 : 0.875749	TRUE		1	2	2:565__647:1005
CAAPA_82741	30	['Spleen : 0.30075290896	7305	233	0.032	7072	['0.0 : 0.695961	TRUE		2	4	2:207__2:306__376:1094
CAAPA_OneEndPlaced_837	26	['Testis : 0.813203300825	2666	66	0.025	2600	['0.0 : 0.518004	FALSE	['exon : ZNF488', 'mRNA : ZNF488']			
CAAPA_OneEndPlaced_1257	28	['Testis : 0.764749813294	5356	108	0.020	5248	['0.0 : 0.561426	FALSE	['exon : LOC105378577', 'mRNA : LOC105378577']			

Figure 3.2 | Sample pipeline output summarizing GTEx alignments to APG contigs. In addition to producing statistics on splicing and transcripts from StringTie2 (if applicable), the report also provides vectors of the tissues and the percentage of reads from each, and a vector with a distribution of the read mapping qualities.

Top potentially transcribed candidate contigs were then examined by hand in IGV to determine if they appeared to have the expected intron/exon structure of a gene. Several contigs were deemed to be reasonable candidates, including CAAPA_OneEndPlaced_1042, a contig which had been placed within the MUC19 gene on chromosome 12 (Figure 3.3), and CAAPA_OneEndPlaced_337, which appears to be a variant of 1042. This top candidate, in fact, has been previously examined in a 2011 study which cloned and characterized the MUC19 gene, and reported as a known alternative splicing/transcript variant (Genbank entry HM801863.1)[137]. However, despite being reported in 2011, the sequence of this transcript is still not included in any version of the human reference genome. These transcribed sequences, as well as CAAPA_TwoEndPlaced_218 within MUC16 (Figure 3.4), are of particular interest as mucin genes are known to be involved in asthma. As the CAAPA cohort consists of nearly half asthma cases, we are further examining whether there might be SNPs or small variants within these novel regions that show a significant association with asthma in the original CAAPA sequences, though

the presence of the contigs themselves are not correlated (refer back to Figure 1.11 for an illustration of SNP discovery in novel sequence).

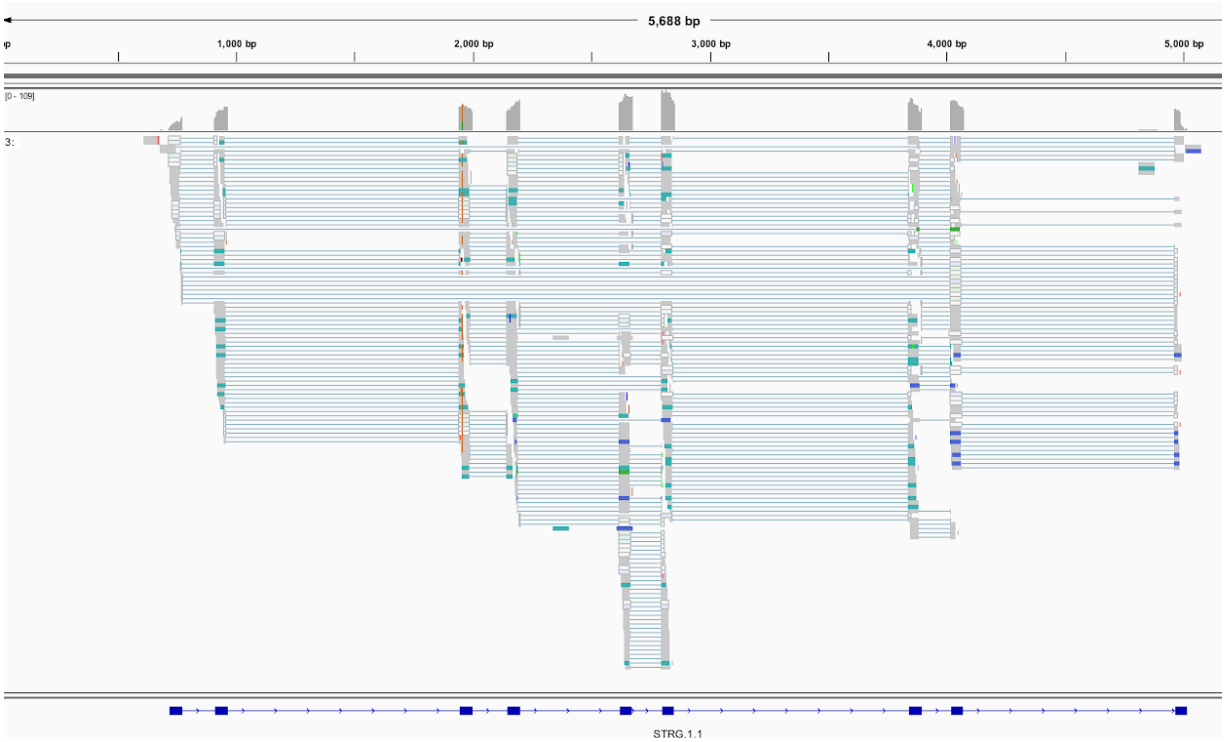


Figure 3.3 | IGV screenshot of spliced GTEx alignments to an APG contig. CAAPA_OneEndPlaced_1042 is 5,688 bases, and appears to contain 9 exons. Nearly all reads aligning to the contig are spliced (indicated by blue lines connecting reads). It has a clear signature of exons, although they are fairly close together, and does not have spurious read alignments in the intronic regions. None of the contig is masked by RepeatMasker.

While not all candidates for novel transcription are as clear candidates as that in Figure 3.3, many candidates without previously reported non-reference sequence also appear to be transcribed. Just based on our preliminary analysis, nearly 30 candidates show promising patterns of spliced alignment, though some are noisier than others (Figure 3.4). We expect using the full GTEx dataset may produce additional candidates, particularly as it includes over 20 additional tissues.

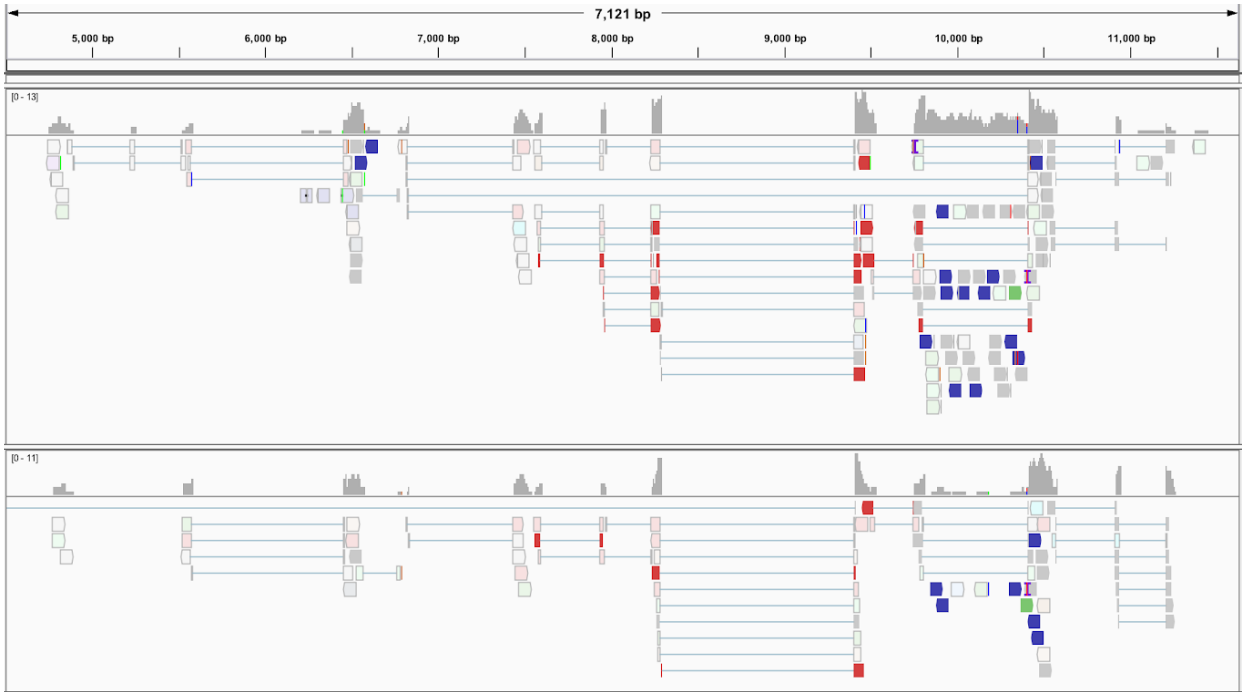


Figure 3.4 | Clearly spliced alignments may still exhibit low coverage and/or noise. Read alignments to this 7,121bp APG contig, CAAPA_TwoEndPlaced_218, placement of which intersect the MUC16 gene, are shown in two distinct samples. Both samples have low (~10x) coverage of the region, but do appear to have spliced alignments indicating exons, though there are un-spliced alignments between two possible exons; possibly due to noise from alignment to repetitive sequence, though this region is not masked by RepeatMasker. A similar alignment pattern is seen in 24 samples across 16 different tissues.

Candidate contigs were additionally run through a translated BLAST search, to look for possible protein homology with other species, even if the sequences are not well conserved at a DNA level. This would provide an additional line of evidence that these human sequences are indeed transcribed. For example, BLAST-X on CAAPA_TwoEndPlaced_218 yields hits to mucin-16 in several primates, covering up to 21% of the contig at over 97% identity (predicted mucin-16-like in Gorilla), providing additional validation that this sequence is placed correctly, transcribed, and part of the human MUC16 gene.

Additional novel sequence sets, many of which are described in Table 1.1, can additionally be added to further extend these analyses. The full v8 GTEx data set can then be aligned to all

available novel sequence sets. As novel sequence sets are now being produced with long reads as well, more sequences are localized in the reference genome, an added benefit to including these in RNA-Seq analyses. A recent study similarly examined alignments of reference-unaligned GTEx reads to their novel sequence set based on novel sequences from 338 human assemblies, finding nearly 5,000 sequences which appeared to be transcribed across all tissues, and additional tissue specific transcription[138]. However, this work too is preliminary, using a subset of the GTEx data -- 10 samples from each of 31 tissues -- and the analysis appears to primarily consider whether reads align, regardless of if they align in a spliced manner. Despite potential shortcomings, though, this analysis again highlights the importance of examining sequences missing from the reference genome; many of these sequences are common, genic, and likely functional, and further analyzing non-reference sequences for transcriptional potential may lead to new insights in functional genomics.

Chapter 4: Utilizing graph-based genotyping to assess disease relevance of structural variants (SVs) detected with long-read sequencing

In this chapter we examine structural variation solely utilizing alignment, rather than assembly based methods, a strategy made possible due to the recent advances in long read (third generation) sequencing technologies. The work in Chapter 4 was primarily overseen by Mike Schatz and portions of Chapter 4 are also described in the following publications, on which I am an author:

Aganezov, S., Goodwin, S., Sherman, R. M., Sedlazeck, F. J., Arun, G., Bhatia, S., ... & Schatz, M. C. (2020). Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Research.*

Chen, S.*, Krusche, P.*, Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., ... & Eberle, M. A. (2019). Paragraph: A graph-based structural variant genotyper for short-read sequence data. *Genome Biology* 20, 291.

This chapter focuses on the aspects of these projects which I was most heavily involved in, regardless of if those portions of the analyses ended up in the final publication. For more complete descriptions of the findings, please refer to the publications themselves.

4.1 Background: Long-read based structural variant detection in breast cancer

Within the field of cancer genomics, dramatic improvements in the throughput and cost of whole-genome sequencing (WGS) and whole-exome sequencing (WES) over the past decade have made these technologies increasingly important in cancer studies, opening the door to widespread sequencing of patients, and the advancement of precision and personalized medicine. Within the Cancer Genome Atlas Project[139], the International Cancer Genome

Consortium[140], the Hartwig Medical Foundation[141], and other large-scale efforts, several thousands of tumors have been sequenced using short-read Illumina sequencing across dozens of major cancer types. These studies have had a tremendous impact in cancer genomics, leading to the discovery, for example, of different signatures and mutation rates across cancer types, and new insights into the clonal structural and evolution of tumors[142–144].

However, despite these advances, we still struggle to identify and understand the genetic alterations in cancer. A major factor contributing to this difficulty is that the known mutations have chiefly been detected using short-read Illumina sequencing[12]. This technology is very effective for identifying single nucleotide variants (SNVs) and large copy number variants (CNVs, especially those 100kb or larger), however, several studies have found it has poor accuracy for structural variant (SV) detection[26]. These SVs, typically defined as variants of 50 bp or larger, where sequence is added, removed, or rearranged in the genome, are difficult to detect with short-read lengths. Short-read Illumina sequencing is difficult to map across SV breakpoints, especially insertions that are not present in the reference genome, where the reads may not map to the reference genome at all (refer to Chapter 1.3 for an overview of SV detection strategies for short reads). As described in Chapter 2, even when we can assemble these insertions from short-reads, they often cannot be subsequently placed back into their genomic context. This is due to the fact that SVs are frequently flanked by repetitive sequences, which means *de novo* assembly techniques fail to capture these novel sequences as well[25]. Consequently, short-read analysis approaches systematically fail to detect SVs, with false negative and false positive rates above 50%[41]. As a result, we are facing a major limitation with short-read sequencing studies of cancer where the field has systematically missed many

important variants, potentially making it blind to entire classes of inherited genetic risk factors and blind to how SVs may mediate cancer progression and patient survival.

New long-read, single molecule sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have been shown to more reliably identify SVs with substantial improvements to both sensitivity and specificity. Reports by several groups have found a typical healthy human genome contains approximately twenty thousand SVs, and that they can be detected with 95% or greater sensitivity and specificity with long-reads[26,27,145]. These variants are especially important to accurately identify for somatic mutations that are not in linkage disequilibrium with any nearby SNVs. Long-reads can also improve the detection of SNVs and smaller insertion/deletion (indel) variants, especially in repetitive sequences and other sequences that are poorly resolved by short-reads[146,147]. Notably, 748 genes have been identified that are inaccessible to short-read sequencing[146], including 193 medically-relevant genes with at least 1 exon that cannot be sequenced with short-reads, but are accessible to long-reads[9,148].

Recent work in cancer genetics led to one of the first reports using PacBio long-read sequencing to study SVs in a cancer cell line genome and found that long-reads could detect tens of thousands of variants that had been missed by short-reads[26,149]. This work examined a Her2 amplified breast cancer cell line, SK-BR-3, detected over 20,000 variants above 50bp in length using long reads. In addition to detecting a vast number of variants, variants in known cancer genes were detected including a complex variant in the HER2 gene, and variants in APOBEC3B and CDH1, as well as dozens of novel gene fusions and other complex rearrangements that had substantially altered the expression and regulation of genes in the cell. This work demonstrated

that these mid-sized variants are of huge import in cancer, that they are quite prevalent, and had gone previously undetected due to the limitations of short reads.

4.2 Structural variant discovery in breast cancer patient organoids

We have performed a similar analysis to the previous SK-BR-3 study on two tumor organoid samples and a matched normal sample from two breast cancer patients, sequenced with Oxford Nanopore, and the resulting analyses are now published in Genome Research[150].

We utilize several tools designed for long read analyses to call structural variants within the samples:

- NGM-LR[26], a long read aligner
- Sniffles[26], a variant caller, designed to work on alignments produced by NGM-LR
- Iris (https://github.com/mkirsche/Iris), which was formerly a module of CrossStitch, (https://github.com/schatzlab/crossstitch) to refine variant call breakpoints and insertion sequences by performing local assemblies around the variant call regions

This pipeline produces a set of variant calls, which can then be filtered by considering the number and percentage of reads supporting the variant, an output of Sniffles. This allows for either filtering of highly confident variants, or only variants which appear to be homozygous. We used this pipeline to analyze PacBio and Oxford Nanopore data from two patient- derived tumor organoid samples, and Nanopore sequencing of normal tissue from the same patient. We additionally examined Illumina (10X Genomics) short-read data to compare the variants which can be detected via long and short reads.

To perform the variant calling, and compare with Illumina short-read variant calls, we utilized an ensemble of methods to infer all types of SVs at least 50bp in size, including insertions, deletions, inversions, translocations, and duplications. For both ONT and PacBio datasets we used two state-of the art methods Sniffles[26] and PBSV (https://github.com/PacificBiosciences/pbsv), and for Illumina/10X dataset we used 6 SV inference methods, with 3 (Lumpy[14], Manta[16], and SvABA[151]) designed for regular paired-end short Illumina reads, and 3 (NAIBR[152], GrocSVS[153], and LongRanger[154]) which also utilize the single-molecule 10X Genomics barcode information. We then iteratively merged SVs using the SURVIVOR[155] package, first merging calls from all SV detection methods for each sequencing technology separately, and then merging across sequencing technologies to obtain sample-specific SV callsets (Figure 4.1a).

Since SVs inferred from paired-end short-reads are notorious for high rates of false positives[19,20,26], for the Illumina/10X dataset we only considered SVs supported by at least 2 methods. To mitigate false positives in the long-read SV calls we only report SVs that were supported by at least one quarter of the average alignment read-depth in either ONT or PacBio datasets. During the merging, we optimize parameters to minimize the effects of small thresholding issues, such as a variant present in 10 reads in one sample, and hence called as a variant, but only 9 reads in other, and hence not called. Our results indicate a very strong concordance between SVs inferred with ONT and PacBio. Between 90% and 95% of variants called in at least one of the long-read data types were supported by both, with slightly lower

concordance between PacBio-only calls (Figure 4.1b). We observe that while more than 50% of SVs inferred from short-read data were also identified by long-reads, the total quantity of SVs inferred from short-reads is approximately 4 times less than for either of the long-read-based inferences.

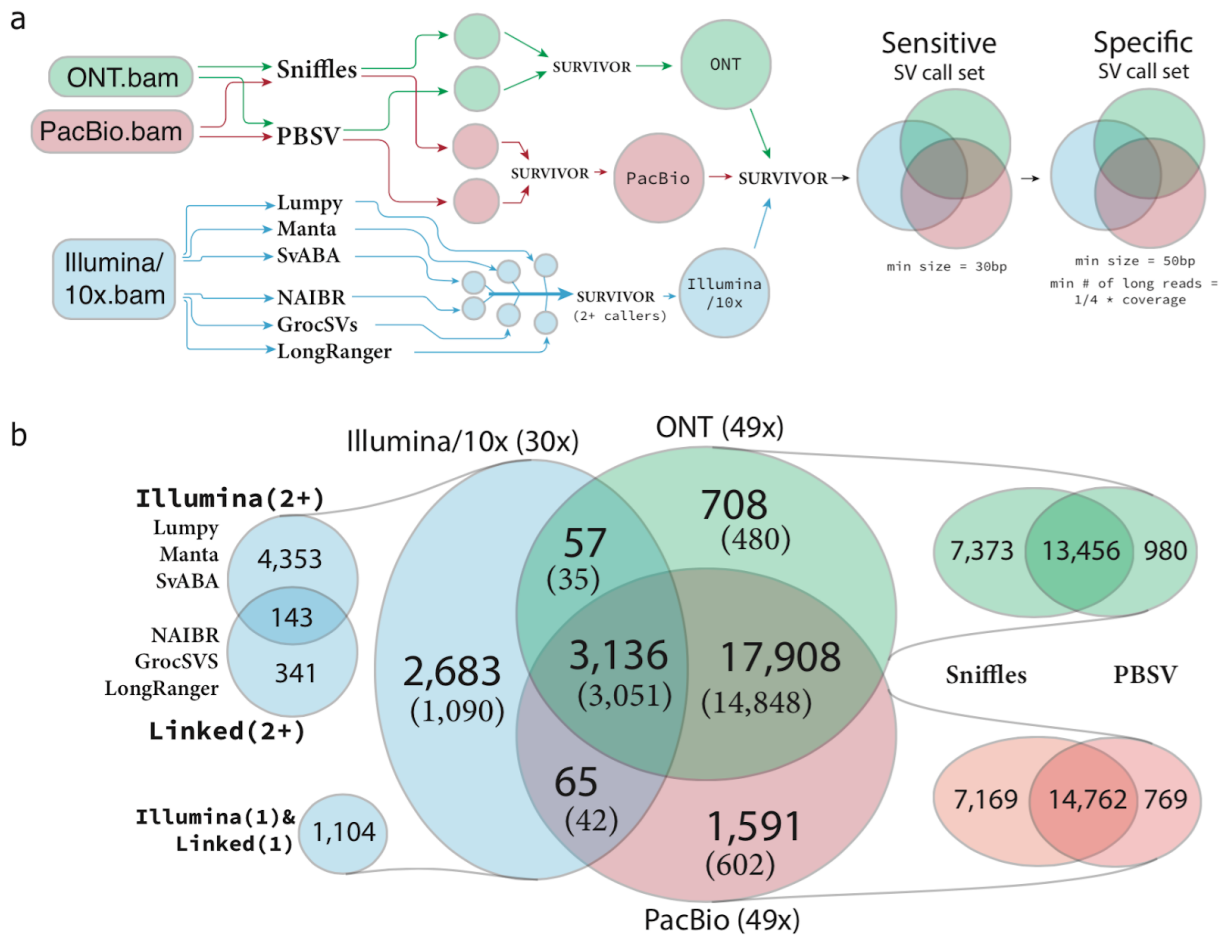


Figure 4.1 | Structural variation inference across sequencing platforms for a patient sample. Patient sample 51 was sequenced with Illumina/10XG, ONT, and PacBio. **a)** Ensemble worfklow for SV inference, with multiple methods and technologies used to infer SVs, subsequent merging of, first method-specific results, and then technology-specific results, with size and support restrictions applied. **b)** SV inference comparison across SVs inferred from *Platform (x)* sequencing experiments, where *Platform* corresponds to sequencing technology, and *(x)* determines the average alignment read-depth coverage in the tumor sample. Methods-specific breakdown is provided for every sequencing technology. SVs detected in the normal sample are in parentheses.

For patient 51 for which we sequenced both the tumor and the matching normal cells we observed that 77% (20,388/26,148) of the SVs identified in the tumor sample were also identified in the matching normal sample (Figure 4.1b). A high fraction of SVs present both in the cancer and in the normal cells is expected since the cancer cells originate from normal tissue. Cancer cells, however, will generally acquire new mutations resulting in the addition of nearly 6,000 variants, although large deletions and loss-of-heterozygosity can potentially decrease the count of inherited SVs[47]. We also observe that for SVs called exclusively by short-reads only ~11% (291/2,683) of SVs inferred in the tumor were also present in the matching normal cells. This is several fold less than for SVs inferred both exclusively with long-reads (88%), and with both long and short-reads (97%), and we attribute this discrepancy to a high false positive rate in short-read SV inference.

To better understand the level of patient-specific and shared germline SVs, both in observed patients and the SK-BR-3 cancer cell-line, we compared SVs inferred with multiple sequencing technologies in the presented study to SVs identified in 15 healthy human genomes sequenced with PacBio long-reads presented in the recent study by Audano *et al*[27]. We find a high level of agreement between the SVs themselves and the distributions of their breakends locations identified in the cancer samples and the healthy samples (Figure 4.2). We observe that 2,577 of the tumor-only SVs in patient 51 are present in other observed healthy samples and we thus hypothesize that many of them are actually present in the normal cells of patient 51, and the inability to infer them in normal cells stems from the lack of coverage in the ONT and the absence of PacBio long-read sequencing of the normal sample. This conjecture is supported by

the comparison of SV types exclusively inferred with different long-read sequencing technologies, since the vast majority (1,806/2,577) are insertions, with ~70% having lengths of 50-200 bp. More accurate basecalling and better SV-genotyping algorithms can help address this limitation in the future.

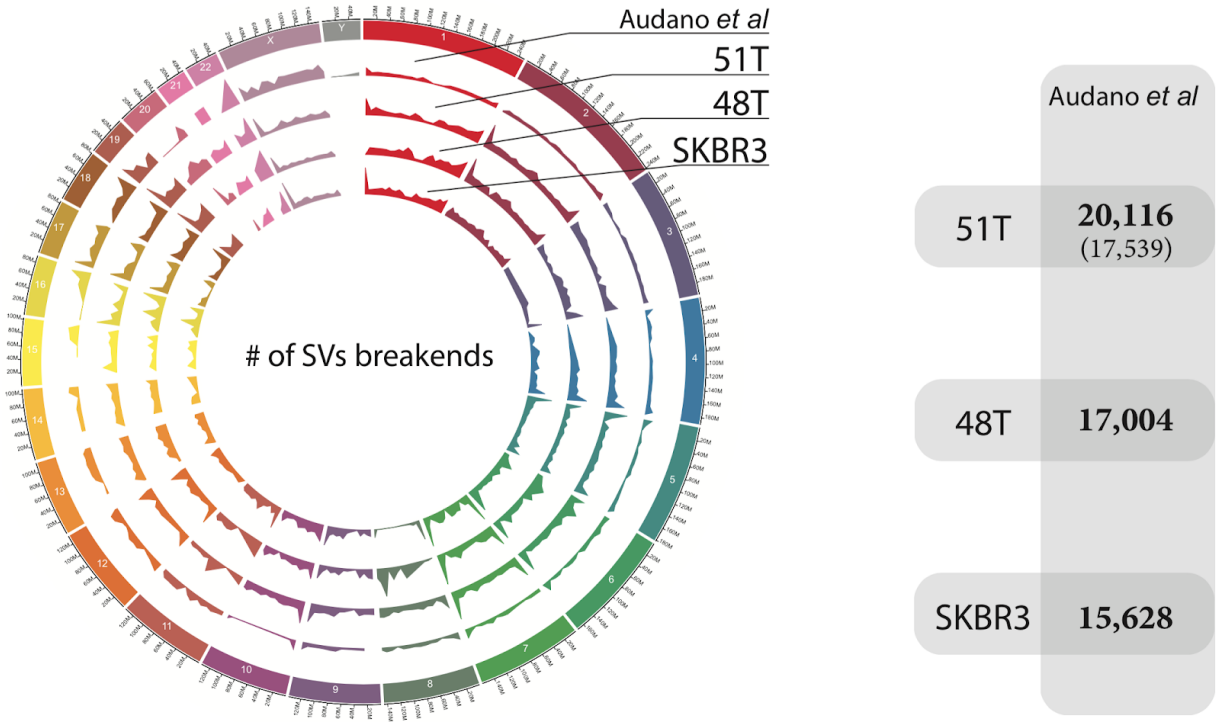


Figure 4.2 | Structural Variations in samples 51T(N), 48T, SK-BR-3, and in Audano *et al*[27] dataset. Circos plot on the left shows the SVs breakends distributions across genome chromosomes. Every track is dataset-specific shows the total number of SVs' breakends over 5MB segment-length windows. Panel on the right shows intersection of SVs across observed cancer datasets (with matching normal SVs shown in parentheses) and the healthy SV set generated from 15 samples from Audano *et al*.

While all variants in the sample were catalogued, we focused further on variants in genes known to be cancer related as part of the COSMIC gene set. On average, more than twice as many SVs (622) affect COSMIC census genes as the genes being affected (237) in 51T. The

majority (199/237) of the SV-affected COSMIC census genes in patient 51 were affected both the tumor and matching normal cells, and furthermore, a majority (466/622) of SVs affecting COSMIC census genes were also present in both the tumor and the matching normal cells. Long-read based SV inference identified five times as many COSMIC census genes affected by SVs and SVs affecting COSMIC census genes than was possible to infer with short-reads. Furthermore, the lack of concordance between SVs inferred exclusively with short-reads between the tumor and normal samples (6/79) provides additional evidence that the short-read SV calling is error-prone. In both patient 48 and the SK-BR-3 cell line we observed similar results with long-read SV inference outperforming short-read SV inference in both the number of COSMIC census genes affected, as well as the number of SVs affecting them. Although these COSMIC affecting SVs may be of relevance to breast cancer, of the 622 SVs, it is likely many are present in healthy individuals as well. To explore this further, short-read WGS datasets are needed, as there are not enough samples available to perform any meaningful analyses of which variants may be cancer associated.

4.3 Development of Paragraph: a short-read structural variant genotyper

While structural variant discovery in long reads is able to discover tens of thousands of novel structural variants, human long-read WGS samples are still being produced slowly. To date, no single human dataset of more than 15 long-read sequenced individuals is readily available, though Iceland has sequenced 1,817 genomes with long reads which are not publicly available[28], and the Human Pangenome Consortium is aiming to sequence 350 genomes, but is still in its pilot phase (https://humanpangenome.org/) of sequencing an initial 10 genomes. The

long-read sequenced samples that are accessible are almost entirely of healthy individuals[27,29]. Meanwhile, a plethora of short-read samples exist for many disease phenotypes including breast cancer, as well as many additional healthy, diverse genomes, such as the 1000 Genomes Project[19], which recently re-sequenced their 2,504 samples with high coverage Illumina data.

While variant discovery in these short-read samples would result in a high false positive rate, if we search for only a specific subset of variants, rather than undertaking a broad search across all locations, this false positive rate can be dramatically reduced. Rather than calling all variants, we can instead aim to genotype variants we suspect may be present. To determine the set of variants to genotype, we leverage the growing number of long-read samples being sequenced to discover structural variants, and then can use a genotyping tool, Paragraph[156], to call these variants in additional samples, and leverage the statistical power of having large cohorts, particularly to determine if a variants are present significantly higher frequencies in disease cohorts vs healthy cohorts or not, an exercise which is impossible with only one of two disease samples when using long reads alone.

We have been actively involved in the development of Paragraph, collaborating with the team at Illumina leading the project. Paragraph uses a graph based method, where the variant is encoded in a graph, and then short reads aligning to the region with the variant are re-aligned to the graph. If they align uniquely to either the graph path with the variant or without the variant, the read is considered for genotyping. To test recall of insertion and deletion genotyping, we calculated the genotyping performance using sequencing data and SVs from the

individual HG002 (also known as NA24385) from Genome in a Bottle (GIAB)[98,157]. We used the short-read sequence data to run Paragraph as well as other methods, and used SVs from long-read sequence data as the ground truth. The short-read data was generated on an Illumina HiSeqX system to 34.5-fold depth using 150bp paired-end reads. The long-read data was generated on a Pacific Biosciences (PacBio) Sequel system using the Circular Consensus Sequencing (CCS) technology[158], to 28-fold coverage with average read length of 13,500 bp. Previous evaluations showed high recall (0.91) and precision (0.94) for SVs called from PacBio CCS HG002 against the GIAB benchmark dataset[98,158], indicating SVs called from CCS data can be effectively used as ground truth to evaluate the performance of SV genotypers and callers. This long-read ground truth (LRGT) set of SVs, all of which are expected to be present in HG002, includes 8,355 deletions and 8,956 insertions. Using this LRGT set, we estimated the performance of Paragraph and two widely-used SV genotypers, SVTyper[159] and Delly Genotyper[15], as well as three methods that independently discover SVs (i.e. *de novo* callers), Manta[16], Lumpy[14] and Delly[15]. Measured against the LRGT calls, Paragraph has the highest recall among all the methods: 0.82 for deletions and 0.82 for insertions (Table 4.1).

Table 4.1 | Recall for different genotypers and de novo callers measured against HG002 LRGT.

Type	Deletion						Insertion	
	Paragraph	Delly Genotyper	SVTyper (100+ bp)	Manta	Lumpy (100+ bp)	Delly (100+ bp)	Paragraph	Manta
# Tested SVs	8,355	8,355	5,372	8,355	5,372	5,372	8,956	8,956
Recall	0.82	0.68	0.35	0.45	0.36	0.21	0.82	0.33
Run time (min)	11	39	16	840*	280	116	14	840*

Genotyping/calling was evaluated using a dataset of HG002 with 150 bp paired-end reads sequenced to 34.5-fold depth on an Illumina HiSeqX. Run time is shown for this data processed on an Intel Xeon E5-2670 2.6GHz eight-core CPUs. Note that SVTyper, Lumpy, and Delly are limited to deletions 100bp or larger so have fewer tested SVs than the other methods.

*Total run time for Manta was 840 minutes for deletions and insertions combined.

However, this test case measuring recall does not represent the primary use-case of Paragraph; to genotype variants discovered in one sample in a different short-read sample. Applying Paragraph to a sample using SVs identified from a large population will also include genotyping variants that are not present in the test sample. To aid in the improvement of the beta version of Paragraph we provided our Illumina collaborators with a test data set where we considered SVs identified from a second sample, ENC002, that was sequenced on PacBio RS II platform as part of the ENCODE project[160].

SVs were called from the PacBio CCS and CLR data using the long read SV caller, Sniffles[26] with parameters "--report-seq -n -1" to report all supporting read names and insertion sequences. Additional default parameters require 10 or more variant supporting reads to report a call, and require variants be at least 50 bp in length. Insertion calls were refined using the insertion refinement module of CrossStitch (https://github.com/schatzlab/crossstitch), which has since been released as a separate program, Iris (https://github.com/mkirsche/Iris) developed by Melanie Kirsche. The module uses FalconSense, an open-source method originally developed for the Falcon assembler[161] and used as the consensus module for Canu[162], to determine a consensus sequence for insertions from variant supporting reads, as Sniffles just reports the sequence taken from a single read. CrossStitch/Iris also performs breakpoint refinement on the variant calls, by aligning the consensus insertion sequence back to the reference region.

Confident reference positions were defined using SVs from CLR ENC002 and CCS HG002. If a deletion is only observed in ENC002 and no deletion is observed 500 bp upstream or

downstream in HG002 with at least 3 supporting reads, this deletion is defined as a confident reference position in HG002. Similarly, if an insertion is only observed in ENC002 and there is no insertion observed in upstream or downstream 200 bp regions in HG002 with at least 3 supporting reads, or there is an insertion in HG002 within 200 bp but their insertion sequences are than 25% concordant, this insertion is defined as a confident reference position in HG002 (Figure 4.3). The precision of Paragraph and other genotypers was then estimated by genotyping these confident reference positions in the short-read HG002. For each genotyper, the recall was calculated as the fraction of SVs in LRGT that were genotyped as non-reference, or the fraction of true positions (TP). The precision was estimated as the fraction of confident reference positions that were genotyped as reference genotypes. The confident reference positions that were genotyped as non-reference are the false positions (FP). Thus, the precision is estimated as TP/(TP+FP).

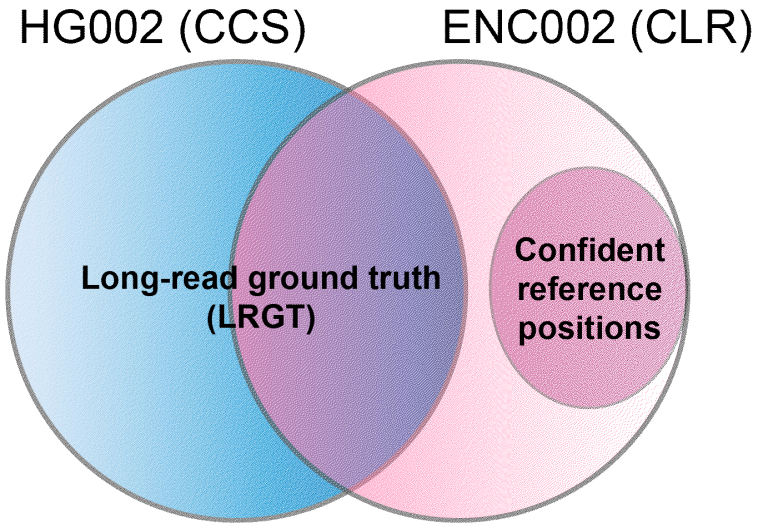


Figure 4.3 | The scheme of building LRGT and confident reference positions. We define SVs called from CCS HG002 as LRGT. We define SVs that were only called in CLR ENC002 and have no same-type SVs called in the nearby region of CCS HG002 as confident reference positions. On short-read HG002, LRGT was used to calculate recall for genotypers and *de novo* callers' recall, while confident reference positions were used to calculate genotypers' precision. Genotypers' precision and F-score were estimated from LRGT and confident reference positions together.

Incorporating the 2,366 deletions and 2,855 insertions that occur in ENC002 but not HG002, the precision for Paragraph was estimated as 0.92 for deletions and 0.90 for insertions (Table 4.2). Notably, the precision for deletions was 0.10 higher than that of Delly Genotyper (0.80). SVTyper is limited to deletions longer than 100bp, and when estimating precision just on the deletions longer than 100bp, Paragraph has a slightly lower precision (0.96) than SVTyper (0.98) though the recall is much higher for Paragraph (0.89 vs 0.35). Combining recall and precision, Paragraph has the highest F-score for deletions in all of the tested genotypers (0.89 vs 0.78 for Delly Genotyper and 0.52 for SVTyper), and also has a high F-score for insertions (0.89). Nearly all (97%) of false positive (FP) deletions and the majority (78%) of FP insertions are completely within TRs. Of the 48 FP insertions that are outside of TRs: 32 have one or more indels (longer than 10 bp) in the target region; 9 have two or more supporting reads for the insertion in HG002 CCS data and those could be false negatives in SV calling from the CCS data; 7 have no evidence of variants in the CCS alignments in the target region, and these FPs likely come from alignment artifacts in short-read mapping.

Table 4.2 | Overall performance for different genotypers.

Type	Deletion				Insertion
	Paragraph	Delly Genotyper	Paragraph (100+ bp)	SVTyper (100+ bp)	Paragraph
#True SVs	8,355		5,372		8,956
#Confident reference positions	2,366		1,001		2,855
Recall	0.82	0.68	0.89	0.35	0.82
Specificity	0.92	0.80	0.96	0.99	0.90
Precision	0.97	0.92	0.98	0.99	0.96
F-score	0.89	0.51	0.93	0.78	0.89

The recall was evaluated on an Illumina HiSeqX sequenced HG002 data using LRGT (same as in Table 4.1). Specificity was evaluated on the same Illumina HiSeqX data using confident reference positions.

4.4 Genotyping variants of interest in large short-read cohorts with Paragraph

Breast cancer patient organoid variant genotyping

To assess the population frequency of the COSMIC affecting breast cancer organoid structural variants, we genotyped identified SVs affecting COSMIC genes from the three analyzed cancer samples with Paragraph[156] in the dataset of 2,504 short-read WGS samples from the recent re-sequencing of the 1000 genomes project (1KGP) samples[19]. Paragraph genotypes SVs by constructing localized sequence graphs containing the reference allele and the candidate SV allele and performs a localized realignment of paired-end short reads to the graph. The genotype is then determined based on the coverage of reads uniquely supporting the reference or variant allele breakpoints. Not all variants can be genotyped by Paragraph in all samples, resulting in no genotype call when support is ambiguous, so we consider only SVs that Paragraph was capable of genotyping in at least 1000 samples.

Table 4.3 | Genotyping of COSMIC gene affecting SVs in 1KGP and Audano *et al* datasets.

Sample	Number of SVs [l|s]	Number of SVs in COSMIC genes [l|s]	1KGP genotyping [l|s]				Not in Audano et al union SV callset & <0.1% in 1KGP
			GT in >1k individuals	< 5%	< 1%	< 0.1%	
51T	26,148 [23,465 | 5,941]	622 [542 | 161]	502 [494|85]	186 [185 | 25]	144 [143 | 17]	112 [111 | 13]	30 [29 | 9]
48T	21,333 [21,333 | NA]	467 [467 | NA]	421 [421|NA]	188 [188 | NA]	156 [156 | NA]	124 [124 | NA]	45 [45 | NA]
SKBR3	20,783 [19,316 | 4,799]	564 [521 | 137]	461 [455|77]	216 [213 |31]	194 [192 | 25]	185 [183 | 23]	121 [119 | 19]

For every observed tumor sample, we report the total number of identified SVs, the number of SVs directly affecting known COSMIC census genes, and the number of COSMIC gene affecting SVs that were successfully genotyped (i.e., called in at least 1000 samples) in 1KGP WGS short-read dataset with frequencies of <5%, <1%, and <0.1 % respectively. For the rarest (i.e., <0.1% in 1KGP) SVs report the number of such SVs that missing in the Audano *et al* union SV set. For every reported SVs count *x* we also show the numbers *[l|s]* of how many of the SVs in *x* were inferred by long *(l)* or short *(s)* reads, respectively.

We then summarize rare variants identified in <5%, <1%, and <0.1% of the overall observed samples (Table 3). We note that Paragraph v2.1 cannot genotype inversions, translocations, and large duplications, and thus we exclude such SVs from the genotyping analysis. SVs that were rarely present in 1KGP individuals (i.e., <0.01% frequency) were further filtered for variants which were not present in any of the 15 healthy genomes from the Audano *et al* study[27]. We show that around 1/5 to 1/4 of the SVs we identified in COSMIC genes are genotyped at low frequency in the 1KGP individuals, and about half of these rarely genotyped SVs are also absent across all of the 15 healthy long-read genomes. These cancer variants found at low-frequency in a healthy population are thus the most likely candidates for cancer risk-factor-type mutations. These variants of interest are identified almost exclusively with long-reads, and although short-read genotyping can help determine population frequency, the ability of 15 long-read samples to additionally narrow the variants of interest further underscores the need for long-read sequenced genomes, both with healthy and disease phenotypes.

Given our candidate set of structural variants which are present in several breast cancer samples, and at low frequency in healthy individuals across populations, the natural question to ask is if these variants are cancer-associated. With only three samples, we are unable to determine if these variants are present at a significantly higher frequency in breast cancer than expected without examining their frequency in a larger breast cancer cohort. The Hartwig Medical Foundation (HMF) and The Cancer Genome Atlas (TCGA) both have hundreds of breast cancer short-read WGS samples available. By using a large breast cancer patient cohort to genotype our discovered LR variants, we will gain the ability to perform statistical analyses

between a large healthy cohort like 1KGP and a large breast cancer cohort. Assuming variants could be found which differ significantly in frequency between the healthy 1KGP cohort and the HMF and/or TCGA breast cancer sample cohort, genes and gene networks can be examined to see if there is a clear potential mechanism that might indicate the variant is causal. RNA-seq data available for a subset of these samples could be utilized to determine if the SVs seem to be having a functional effect, as well. This area of future work will be useful for gaining more insight into association of SVs, rather than just discovering and narrowing candidates of interest, as we have done using the breast cancer organoid samples. These techniques could of course be extended to other cancer types as well, so a small number of LR-sequenced cancer samples can be utilized in conjunction with Paragraph and pre-existing short-read cancer and healthy samples.

Rare Mendelian disorder variant discovery and genotyping

Another area in which we are utilizing Paragraph is in an examination of several families with rare Mendelian disorders. Two families appear to have X-linked disorders, based on the pattern of inheritance in the pedigrees (Figure 4.4). Affected individuals in family 1 experience a severe slowly degenerative muscular myopathy, and six individuals have been sequenced with ONT reads, four affected and two unaffected males. The depth of sequencing coverage varies, but is approximately 20-30x per individual. Family 2 affected individuals experience a cleft palate and underdeveloped facial structure phenotype. As affected individuals in family 2 do not often make it past childhood, sequencing data is more limited. We have ONT data from one affected individual and his mother, a presumed carrier. An additional sample has been secured from one

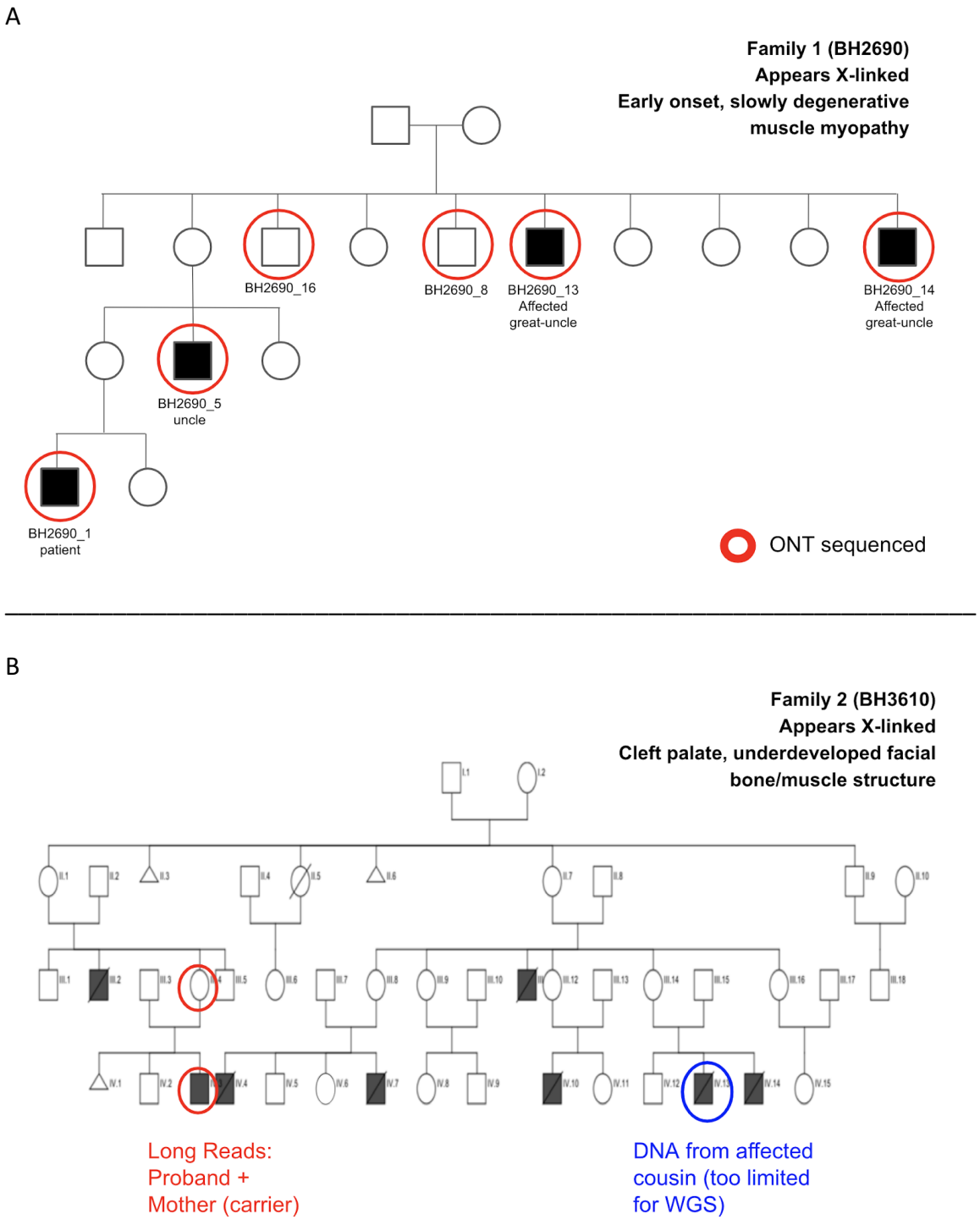
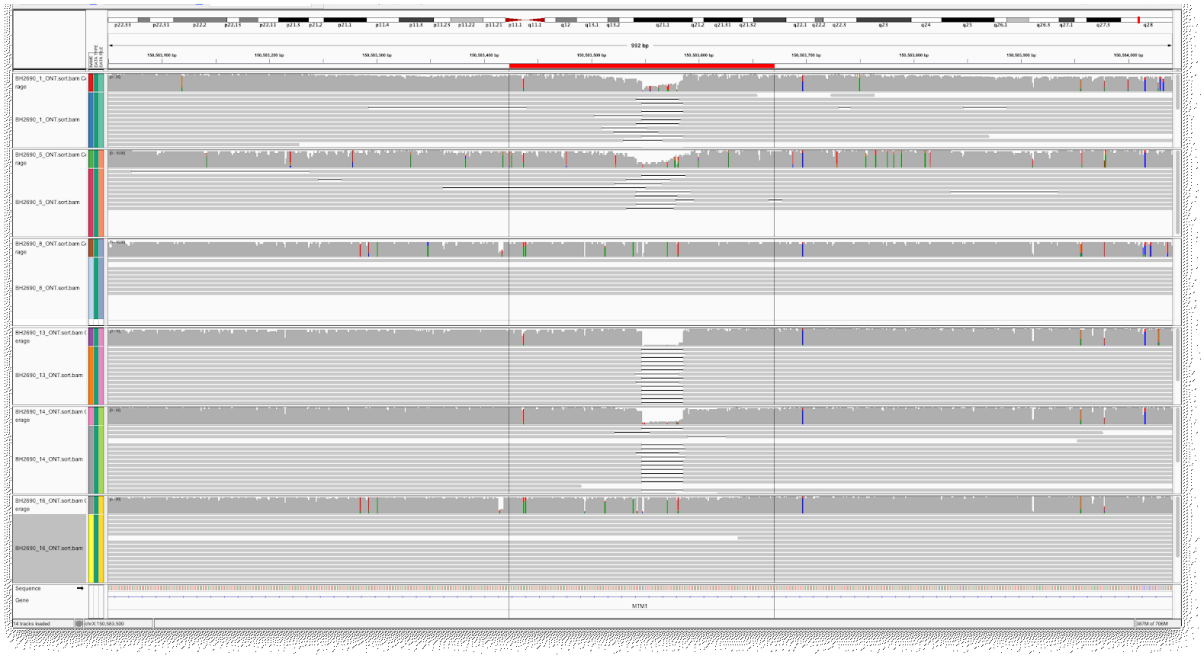


Figure 4.4 | Pedigrees of two ONT-sequenced families with rare mendelian disorders. (A) Family 1 has a slowly degenerative muscular myopathy that appears to be X-linked, as it only affects males. Six males in the family, four affected, and two unaffected, have been sequenced with ONT to ~20-30x coverage. (B) Family 2 exhibits a cleft palate disorder, and also appears to have an X-linked inheritance pattern. An unaffected (carrier) mother and her affected son have been sequenced with long reads. Additional DNA from an affected male is available to perform PCR validations, but not sequencing.

other affected individual in the family, however, there is not enough sample to sequence -- it is being reserved for validation experiments, if needed.

Alignment with NGM-LR and variant calling with Sniffles[26] was performed on the sequenced individuals, using the same methods described in Chapter 4.2 and Aganezov *et al* 2020[150] to produce both sensitive and specific call sets. SVs of individuals within families were then compared to find SVs common to affected individuals but absent from unaffected individuals. Once these SVs were identified, they were further screened against both individuals in the other family, and the SVs called from Audano *et al* 2019[27], initially using SURVIVOR[163] (as described in Chapter 4.2). In this family-based analysis, where very few variants are expected between family members, especially the mother and son in Family 2, it quickly became evident that small issues with SURVIVOR's merging algorithm were presenting problems; variants were both being merged erroneously (for example, when variants were different sizes but in nearby locations) or erroneously left unmerged. To mitigate this, the merging tool Jasmine has been developed, led by Melanie Kirche (https://github.com/mkirsche/Jasmine, paper forthcoming). Switching from SURVIVOR to Jasmine produced additional candidate variants, such as a 35 bp intronic deletion in Family 1 in the MTM1 gene on the X chromosome, a gene with other variants known to be related to myopathy. This variant had not initially passed screening, because a 450 bp deletion exists in other, non-family 1 long read samples that overlaps the smaller deletion's location (Figure 4.5). Having a merging tool such as Jasmine, which accounts for not just location, but size and type of a variant, is crucial to uncovering potentially disease related variants, as candidates such as these are missed using other merging software.

A



B

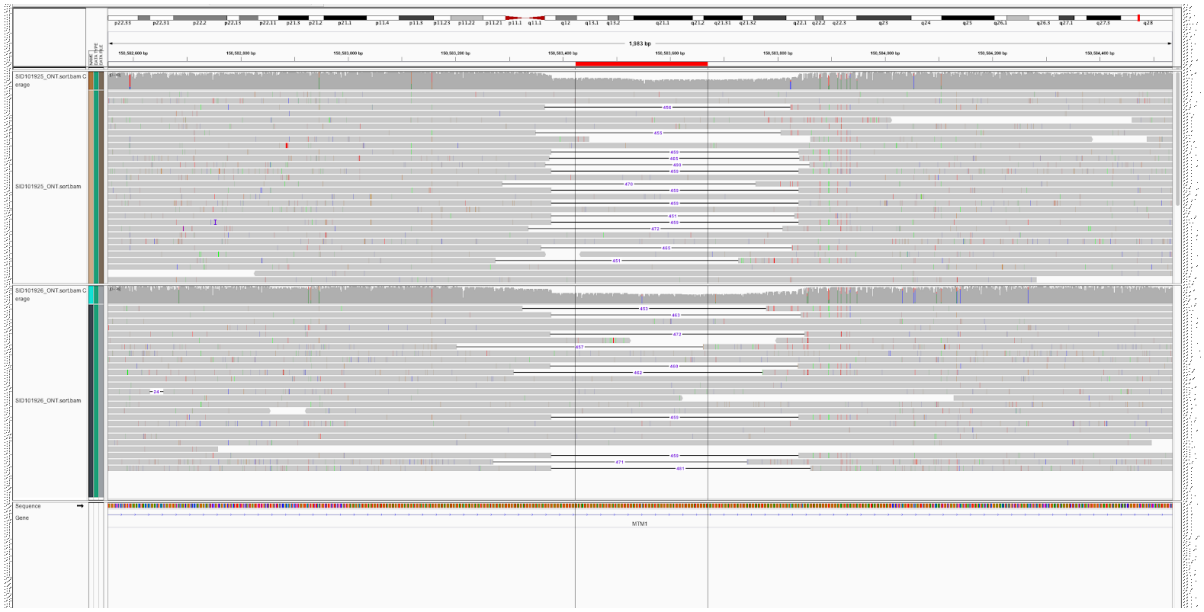


Figure 4.5 | MTM1 deletion candidate overlaps larger deletion in healthy individuals. (A) A ~35 bp deletion is present in the four affected individuals (rows 1,2,4,5) in Family 1 and in neither of the unaffected individuals (rows 3,6). (B) A ~450 bp deletion overlapping the coordinates is present in several healthy individuals (matching coordinates indicated by vertical lines in (A) and (B). This demonstrates the need for careful merging of SVs in filtering analyses.

We then used Paragraph to screen the candidate X chromosome variants from both families in the 1KGP cohort, using the same method as for the breast cancer patient structural variants. Although several candidate variants in each family had low frequency in the 1KGP cohort, and were present only in affected individuals, thus far no candidate variant has segregated as expected when examined by PCR amplification in the families, where individuals without long read sequencing data were included in the PCR based validation. (These analyses were performed by our collaborator, Nara Sobreira, and members of her lab, who also recruited the families in the study and have procured samples as needed.) This work is ongoing, and we hope that as we continue to fine-tune and improve our variant calling, merging, and large cohort screening pipeline, that we might discover a disease-associated variant either in these or additional families. As Mendelian disease associated variants have been discovered recently using long read sequencing[92,164–166], we are hopeful that overall this strategy will prove to be effective, even if a large SV is not ultimately the culprit in these two particular families.

Conclusions

For many years scientists have understood that considering more than just a single representative genome can help identify genes and phenotypically consequential variants in bacterial and plant species. However, in human studies, nearly all analyses still begin by aligning sequence data from a subject or a set of subjects to the human reference genome, discarding sequences that do not align. Considering more than just a single reference genome is necessary if we are to link more phenotypes of interest to their causal variants. From the work presented in this thesis as well as from other recent studies, we now know that populations across the globe contain many thousands of DNA sequences that are not present in the human reference genome and thus not examined in standard analyses. Although we are amassing a wealth of pan-genomic data in both global and population-specific studies, what to do with these data remains an open question. The creation of a single, global human pan-genome holds conceptual appeal, and cataloguing all human variation is a noble goal. However, to date no computational method is capable of aligning human sequences to a pan-genome of all human variation, while enforcing that alignments be biologically plausible, and without introducing additional alignment ambiguities, although these subjects are all active research areas.

Population-specific pan-genomes may prove more feasible, and multiple such projects such as the Icelandic and Danish efforts are underway. Developing additional linear reference genomes has the benefit of representing a real individual, and a linear representation does not introduce variants that are never seen together. Furthermore, our ability to accurately assemble linear reference genomes is rapidly improving: recently the first telomere-to-telomere assembly of a

human chromosome[125] demonstrated that newly produced genomes have the potential to be of much higher contiguity than the current reference, and the Telomere-to-Telomere Consortium has since finished additional human chromosomes. Long read sequencing has been critical to these efforts, and although long read sequencing is still not the standard, it is rapidly growing in use and declining in cost. In addition to providing better assembly resolution, long reads provide previously unprecedented resolution for structural variant detection, even when using a reference-based approach.

As we have demonstrated in our analyses of several cancer organoid samples, robust SV detection is possible at relatively low ~30x average read-depth coverage with either ONT or PacBio long read sequencing platforms. When applied at scale, costs for 30x coverage is below $1000 per sample for ONT PromethION and below $2,000 for PacBio CLR Sequel II, which is highly comparable to ~$800/$1,000 (Illumina/10XG) for short read sequencing, a good indication that long read sequencing is the way forward. However, the lower throughput nature of single molecule sequencing, and the fact that new technologies are often slow to be approved and adopted for clinical use, makes hybrid approaches utilizing available short read and a small amount of long read data, appealing in the interim. Utilizing graph-based genotyping with Paragraph, even a small number of long read sequenced samples are a valuable resource to examine structural variants of interest at a population scale. In our examination of several breast cancer patient samples, we initially discovered thousands of novel structural variants detected via long read sequencing, and genotyping in a healthy cohort enabled this variant set to be narrowed significantly. While we wait for ONT and PacBio

technologies to become more widely adopted and for the creation of new, long read datasets, we must utilize the data already available to make clinically actionable discoveries, now.

Eventually we will be armed with a plethora of long read derived, highly contiguous, near-perfect quality, human reference genomes, and individuals can be analyzed by comparing them to their closest matching population or populations, even if this information is not known a priori. And although the representations we will ultimately use to align to these genomes while taking known variation into account is still unclear, it seems inevitable that we will soon move beyond our reliance on a single human reference genome. We have shown here that hundreds of megabases of novel sequence can be found in a modestly sized cohort, that any given individual has on the order of 20,000 structural variants from the reference genome which can only be reliably detected with long reads, and that many of these variants are not just within genes, but in genes with known disease relevance. Finding a way to capture and consider these vast amounts of variation in the population will be critical moving forward if we are ultimately to truly understand the genetic instructions that make us human.

References

1. Anon. Human Genome Project FAQ. 2019.
2. The International Human Genome Sequencing Consortium, Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, et al. The sequence of the human genome. *Science*. 2001;291(5507):1304–1351.
4. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*. 2017;27(5):849–864.
5. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, et al. A draft sequence of the Neandertal genome. *Science*. 2010;328(5979):710–722.
6. Illumina Inc. An Introduction to Next-Generation Sequencing Technology. *Illumina*. 2017.
7. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53–59.
8. Sipos B, Massingham T, Stütz AM, Goldman N. An improved protocol for sequencing of

repetitive genomic regions and structural variations using mutagenesis and next generation sequencing. *PloS one*. 2012;7(8):e43359.

9. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*. 2019;37(10):1155–1162.
10. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nature reviews. Genetics*. 2020;21(10):597–614.
11. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*. 2020;21(1):30.
12. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews. Genetics*. 2016;17(6):333–351.
13. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. *Briefings in functional genomics*. 2015;14(5):305–314.
14. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome biology*. 2014;15(6):R84.
15. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* . 2012;28(18):i333–i339.
16. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and

cancer sequencing applications. *Bioinformatics* . 2016;32(8):1220–1222.

17. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT. Strelka2: fast and accurate calling of germline and somatic variants. *Nature methods*. 2018;15(8):591–594.
18. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]*. 2012.
19. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75–81.
20. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin C-S, Korlach J, Wilson RK, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*. 2017;27(5):677–685.
21. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MH, Cao H, Cohain A, Deikus G, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*. 2015;12(8):780–786.
22. Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HYK. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* . 2015;31(16):2741–2744.
23. English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S, Carroll A, Veeraraghavan N, Bruestle J, Drees B, Hastie A, et al.

Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC genomics*. 2015;16:286.

24. Zarate S, Carroll A, Krashenina O, Sedlazeck FJ, Jun G, Salerno W, Boerwinkle E, Gibbs R. Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers. *Cold Spring Harbor Laboratory*. 2018:424267.
25. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*. 2015;517(7536):608–611.
26. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*. 2018;15(6):461–468.
27. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*. 2019;176(3):663–675 e19.
28. Beyter D, Ingimundardottir H, Eggertsson HP, Bjornsson E, Kristmundsdottir S, Mehringer S, Jonsson H, Hardarson MT, Magnusdottir DN, Kristjansson RP, Gudjonsson SA, Sverrisson ST, Holley G, Eyjolfsson G, Olafsson I, et al. Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. *bioRxiv*. 2019:848366.
29. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, Sedlazeck FJ, Marschall T, Mayes S, Costa V, Zook JM, et al. Efficient de

novo assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit. *bioRxiv*. 2019:715722.

30. Rouli L, Merhej V, Fournier PE, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect*. 2015;7:72–85.
31. Pallen MJ, Wren BW. Bacterial pathogenomics. *Nature*. 2007;449(7164):835–842.
32. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(39):13950–13955.
33. Ali A, Naz A, Soares SC, Bakhtiar M, Tiwari S, Hassan SS, Hanan F, Ramos R, Pereira U, Barh D, Figueiredo HC, Ussery DW, Miyoshi A, Silva A, Azevedo V. Pan-genome analysis of human gastric pathogen H. pylori: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. *BioMed research international*. 2015;2015:139580.
34. Ali A, Soares SC, Santos AR, Guimaraes LC, Barbosa E, Almeida SS, Abreu VA, Carneiro AR, Ramos RT, Bakhtiar SM, Hassan SS, Ussery DW, On S, Silva A, Schneider MP, et al. Campylobacter fetus subspecies: comparative genomics and prediction of potential virulence targets. *Gene*. 2012;508(2):145–156.
35. Imperi F, Antunes LCS, Blom J, Villa L, Iacono M, Visca P, Carattoli A. The genomics of Acinetobacter baumannii: insights into genome plasticity, antimicrobial resistance and pathogenicity. *IUBMB life*. 2011;63(12):1068–1074.

36. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of bacteriology*. 2008;190(20):6881–6893.

37. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. *Genome research*. 2015;25(1):119–128.

38. Trost E, Blom J, Soares Sde C, Huang IH, Al-Dilaimi A, Schroder J, Jaenicke S, Dorella FA, Rocha FS, Miyoshi A, Azevedo V, Schneider MP, Silva A, Camello TC, Sabbadini PS, et al. Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. *Journal of bacteriology*. 2012;194(12):3199–3215.

39. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. *Current opinion in genetics & development*. 2005;15(6):589–594.

40. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65.

41. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature reviews. Genetics*. 2018;19(6):329–346.

42. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome research*. 2014;24(4):697–707.

43. Kehr B, Helgadottir A, Melsted P, Jonsson H, Helgason H, Jonasdottir A, Jonasdottir A,

Sigurdsson A, Gylfason A, Halldorsson GH, Kristmundsdottir S, Thorgeirsson G, Olafsson I, Holm H, Thorsteinsdottir U, et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nature genetics*. 2017;49(4):588–593.

44. Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H, Gylfason A, et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci Data*. 2017;4:170115.
45. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, Skov L, Belling K, Theil Have C, Izarzugaza JMG, Grosjean M, Bork-Jensen J, Grove J, Als TD, Huang S, et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*. 2017;548(7665):87–91.
46. Eisfeldt J, Mårtensson G, Ameur A, Nilsson D, Lindstrand A. Discovery of Novel Sequences in 1,000 Swedish Genomes. *Molecular biology and evolution*. 2020;37(1):18–30.
47. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, Mondal M, Pagani L, Ricaut F-X, Stoneking M, Metspalu M, Sudoyo H, Lansing JS, Cox MP. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*. 2019;177(4):1010–1021.e32.
48. Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, Zhang Y, Bond SR, Pei Z, Zhang Y, Zhang D, Jirimutu J, Zhang D, Yang X, Morigenbatu M, et al. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nature genetics*. 2018;50(12):1696–1704.
49. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, Chimusa ER, Christoffels A, Gamieldien J, Sefid-Dashti MJ, Joubert F, Meintjes A, Mulder N, Ramesar R, Rees J, et al.

Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature communications*. 2017;8(1):2062.

50. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GR, Xue Y, Asimit J, Nsubuga RN, Young EH, et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature*. 2015;517(7534):327–332.
51. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, Vergara C, Torgerson DG, Pino-Yanes M, Shringarpure SS, Huang L, Rafaels N, Boorgula MP, Johnston HR, Ortega VE, et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nature communications*. 2016;7:12522.
52. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland N, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016;538(7624):201–206.
53. Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C, Brewerton SC, Bulsara N, Garner C, Metzker G, Sandoval E, et al. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*. 2016;113(42):11901–11906.
54. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
55. Hall SS. Revolution postponed. *Scientific American*. 2010;303(4):60–67.

56. Wade N. A decade later, genetic map yields few new cures. *The New York times*. 2010;12.

57. Francis WR, Wörheide G. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. *Genome biology and evolution*. 2017;9(6):1582–1598.

58. Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. Human protein-coding genes and gene feature statistics in 2019. *BMC research notes*. 2019;12(1):315.

59. Ganguly P. NHGRI funds centers for advancing the reference sequence of the human genome. 2019.

60. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*. 2001;29(1):308–311.

61. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*. 2014;42(Database issue):D980–5.

62. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*. 2002;30(1):52–55.

63. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*. 2019;37(8):907–915.

64. Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic acids research*. 2014;42(Web Server issue):W494–500.

65. Tithi SS, Heath LS, Zhang L. *SNPwise: A SNP-aware short read aligner*.; 2015.

66. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, Flicek P, Church DM. DbVar and DGVa: public archives for genomic structural variation. *Nucleic acids research*. 2013;41(Database issue):D936–41.

67. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic acids research*. 2014;42(Database issue):D986–92.

68. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Gagliano Taliun SA, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee S-B, Tian X, Browning BL, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*. 2019:563866.

69. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75–81.

70. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome biology*. 2019;20(1):92.

71. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, Madugundu AK, Pandey A, Salzberg SL. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome biology*. 2018;19(1):208.

72. Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, Renkens I, Coe BP, Deelen P, de Ligt J, Lameijer E-W, et

al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature communications*. 2016;7:12989.

73. Miga KH. Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes*. 2019;10(5).
74. Barra V, Fachinetti D. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nature communications*. 2018;9(1):4340.
75. Ameur A, Che H, Martin M, Bunikis I, Dahlberg J, Hoijer I, Haggqvist S, Vezzi F, Nordlund J, Olason P, Feuk L, Gyllensten U. De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. *Genes*. 2018;9(10).
76. Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, Sun C, Hu Z, Zhang Z, Li G, Chen H, Xiang Z, Zhu Z, Zhao H, Yu Y, et al. HUPAN: a pan-genome analysis pipeline for human genomes. *Genome biology*. 2019;20(1):149.
77. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, Chung CYL, Ma W, Sibert J, Rajagopalan R, Jin N, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature communications*. 2019;10(1):1025.
78. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, Levin AM, Eng C, Yazdanbakhsh M, Wilson JG, Marrugo J, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature genetics*. 2019;51(1):30–35.

79. Wong KHY, Levy-Sakin M, Kwok P-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature communications*. 2018;9(1):3040.

80. Faber-Hammond JJ, Brown KH. Anchored pseudo-de novo assembly of human genomes identifies extensive sequence variation from unmapped sequence reads. *Human genetics*. 2016;135(7):727–740.

81. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, Sato Y, Mimori T, Tsuda K, Saito R, Pan X, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature communications*. 2015;6:8018.

82. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, et al. Modernizing reference genome assemblies. *PLoS biology*. 2011;9(7):e1001091.

83. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin CS, Kitts PA, Aken B, Marth GT, Hoffman MM, Herrero J, Mendoza ML, Durbin R, Flicek P. Extending reference assembly models. *Genome biology*. 2015;16:13.

84. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* . 2009;25(14):1754–1760.

85. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. *Nature reviews. Genetics*. 2020.

86. Computational Pan-Genomics Consortium,. Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*. 2018;19(1):118–135.

87. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome

inference. *Genome research*. 2017;27(5):665–676.

88. Pritt J, Chen N-C, Langmead B. FORGe: prioritizing variants for graph genomes. *Genome biology*. 2018;19(1):220.
89. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*. 2018;36(9):875–879.
90. Rakocevic G, Semenyuk V, Lee W-P, Spencer J, Browning J, Johnson IJ, Arsenijevic V, Nadj J, Ghose K, Suciu MC, Ji S-G, Demir G, Li L, Toptaş BÇ, Dolgoborodov A, et al. Fast and accurate genomic analyses using genome graphs. *Nature genetics*. 2019;51(2):354–362.
91. Grytten I, Rand KD, Nederbragt AJ, Sandve GK. Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC genomics*. 2020;21(1):282.
92. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, Kohno Y, Takiyama Y, Maeda K, Doi H, Koyano S, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nature genetics*. 2019;51(8):1215–1221.
93. Anon. E pluribus unum. *Nature methods*. 2010;7(5):331.
94. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. *Trends in genetics: TIG*. 2009;25(11):489–494.
95. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161–164.

96. Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, Zhou G, Zhu X, Wu H, Qin J, Jin X, et al. Building the sequence map of the human pan-genome. *Nature biotechnology*. 2010;28(1):57–63.

97. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J, Kuk J, Park GH, Kim J, Ryu H, Kim J, et al. De novo assembly and phasing of a Korean human genome. *Nature*. 2016;538(7624):243–247.

98. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre AB, Chandramohan D, Chen F, Jaeger E, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025.

99. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, Lintner KE, Ding Q, Wang Z, Hu J, Wang D, et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nature communications*. 2016;7:12065.

100. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, Chae KS, Kim CG, Kim S, Eriksson A, Edwards JS, Lee S, Kim BC, Manica A, Oh TK, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nature communications*. 2016;7:13637.

101. Shumate A, Zimin AV, Sherman RM, Puiu D, Wagner JM, Olson ND, Pertea M, Salit ML, Zook JM, Salzberg SL. Assembly and annotation of an Ashkenazi human reference genome. *Genome biology*. 2020;21(1):129.

102. Kehr B, Melsted P, Halldorsson BV. PopIns: population-scale detection of novel sequence insertions. *Bioinformatics* . 2016;32(7):961–967.

103. Gordienko EN, Kazanov MD, Gelfand MS. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *Journal of bacteriology*. 2013;195(12):2786–2792.

104. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Current opinion in microbiology*. 2015;23:148–154.

105. Crysnanto D, Pausch H. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome biology*. 2020;21(1):184.

106. Li R, Fu W, Su R, Tian X, Du D, Zhao Y, Zheng Z, Chen Q, Gao S, Cai Y, Wang X, Li J, Jiang Y. Towards the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the Reference Genome. *Frontiers in genetics*. 2019;10:1169.

107. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, Wang Y, Fan D, Zhao Y, Wang Z, Zhou C, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature genetics*. 2018;50(2):278–284.

108. Sun C, Hu Z, Zheng T, Lu K, Zhao Y, Wang W, Shi J, Wang C, Lu J, Zhang D, Li Z, Wei C. RPAN: rice pan-genome browser for approximately 3000 rice genomes. *Nucleic acids research*. 2017;45(2):597–605.

109. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics*. 2019;51(6):1044–1051.

110. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang SS, Zuo Q, Shi XH, Li YF, Zhang WK, et al. De novo assembly of soybean wild relatives for pan-genome

analysis of diversity and agronomic traits. *Nature biotechnology*. 2014;32(10):1045–1052.

111. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA, Paterson AH, Pires JC, Sharpe AG, Tang H, Teakle GR, et al. The pangenome of an agronomically important crop plant Brassica oleracea. *Nature communications*. 2016;7:13390.
112. Hubner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, Lee JS, Baute GJ, Owens GL, Grassa CJ, Ebert DP, Ostevik KL, Moyers BT, Yakimowski S, Masalia RR, et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat Plants*. 2019;5(1):54–62.
113. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012;9(4):357–359.
114. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* . 2009;25(16):2078–2079.
115. Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. *Bioinformatics* . 2013;29(21):2669–2677.
116. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*. 2016;26(12):1721–1729.
117. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*. 2009;Chapter 4:Unit 4 10.
118. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large

sequence sets. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*. 2003;Chapter 10:Unit 10 3.

119. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* . 2010;26(6):841–842.
120. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*. 2014;15(3):R46.
121. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC bioinformatics*. 2009;10:421.
122. Siren J, Garrison E, Novak AM, Paten B, Durbin R. Haplotype-aware graph indexes. *Bioinformatics* . 2019.
123. Kuhnle A, Mun T, Boucher C, Gagie T, Langmead B, Manzini G. *Efficient Construction of a Complete Index for Pan-Genomics Read Alignment*. Cham: Springer International Publishing; 2019.
124. Gagie T, Navarro G, Prezza N. Optimal-time text indexing in BWT-runs bounded space. In: Society for Industrial and Applied Mathematics; 2018:1459–1477.
125. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, Schneider VA, Potapova T, Wood J, Chow W, Armstrong J, et al. Telomere-to-telomere assembly of a complete human X chromosome. *bioRxiv*. 2019:735928.
126. Soifer L, Fong NL, Yi N, Ireland AT, Lam I, Sooknah M, Paw JS, Peluso P, Concepcion GT, Rank D, Hastie AR, Jojic V, Ruby JG, Botstein D, Roy MA. Fully Phased Sequence of a Diploid Human Genome Determined de Novo from the DNA of a Single Individual. *G3* .

2020;10(9):2911–2925.

127. Manni M, Zdobnov E. Microbial contaminants cataloged as novel human sequences in recent human pan-genomes. 2020:2020.03.16.994376.

128. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature communications*. 2016;7:11257.

129. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature methods*. 2015;12(1):59–60.

130. Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome biology*. 2020;21(1):115.

131. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* . 2013;29(1):15–21.

132. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*. 2013;14(4):R36.

133. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*. 2017;35(12):1167–1169.

134. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome biology*. 2019;20(1):278.

135. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318–1330.

136. Sá ACC, Sadee W, Johnson JA. Whole Transcriptome Profiling: An RNA-Seq Primer and Implications for Pharmacogenomics Research. *Clinical and translational science*.

2018;11(2):153–161.

137. Zhu L, Lee P, Yu D, Tao S, Chen Y. Cloning and characterization of human MUC19 gene. *American journal of respiratory cell and molecular biology*. 2011;45(2):348–358.

138. Wong KHY, Ma W, Wei C-Y, Yeh E-C, Lin W-J, Wang EHF, Su J-P, Hsieh F-J, Kao H-J, Chen H-H, Chow SK, Young E, Chu C, Poon A, Yang C-F, et al. Towards a reference genome that captures global genetic diversity. *Nature communications*. 2020;11(1):5482.

139. Weinstein JN, The Cancer Genome Atlas Research Network, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. 2013;45(10):1113–1120.

140. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, Wong-Erasmus M, Yao L, Kasprzyk A. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database: the journal of biological databases and curation*. 2011;2011:bar026.

141. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, Duyvesteyn K, Haidari S, van Hoeck A, Onstenk W, Roepman P, Voda M, Bloemendal HJ, Tjan-Heijnen VCG, van Herpen CML, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*. 2019;575(7781):210–216.

142. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. *Nature reviews. Genetics*. 2017;18(4):213–229.

143. Yates LR, Campbell PJ. Evolution of the cancer genome. *Nature reviews. Genetics*. 2012;13(11):795–806.

144. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli

N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.

145. De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M, Sleegers K, Van Broeckhoven C. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome research*. 2019;29(7):1178–1187.
146. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, Kauwe JSK, Belzil V, Pregent L, Carrasquillo MM, Keene D, Larson E, Crane P, Asmann YW, Ertekin-Taner N, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome biology*. 2019;20(1):97.
147. Luo R, Sedlazeck FJ, Lam T-W, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nature communications*. 2019;10(1):998.
148. Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, Duffy E, Hegde M, Santani A, Lebo M, Funke B. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genetics in medicine: official journal of the American College of Medical Genetics*. 2016;18(12):1282–1289.
149. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin T, Fang H, Gurtowski J, Hutton E, Tseng E, Chin C-S, Beck T, Sundaravadanam Y, Kramer M, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome research*. 2018;28(8):1126–1135.
150. Aganezov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, Bhatia S, Lee I, Kirsche M, Wappel R, Kramer M, Kostroff K, Spector DL, Timp W, McCombie WR, Schatz MC. Comprehensive analysis of structural variants in breast cancer genomes using

single-molecule sequencing. *Genome research*. 2020;30(9):1258–1273.

151. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X, Nusbaum C, Campbell P, Getz G, Meyerson M, Zhang C-Z, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome research*. 2018;28(4):581–591.
152. Elyanow R, Wu H-T, Raphael BJ. Identifying structural variants using linked-read sequencing data. *Bioinformatics* . 2018;34(2):353–360.
153. Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, West RB, Batzoglou S, Sidow A. Genome-wide reconstruction of complex structural variants using read clouds. *Nature methods*. 2017;14(9):915–920.
154. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA, Wyatt PW, Bharadwaj R, Makarewicz AJ, Li Y, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*. 2016;34(3):303–311.
155. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications*. 2017;8:14061.
156. Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ, Eberle MA. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome biology*. 2019;20(1):291.
157. Zook J, McDaniel J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, De La Vega FM, Xiao C, Sherry S, Salit M, Genome in a Bottle Consortium. Reproducible integration of

multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials.

158. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Toepfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, et al. Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *bioRxiv*. 2019:519025.
159. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature methods*. 2015;12(10):966–968.
160. The ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57.
161. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*. 2016;13(12):1050–1054.
162. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*. 2017;27(5):722–736.
163. Sedlazeck FJ, Dhroso A, Bodian DL, Paschall J, Hermes F, Zook JM. Tools for annotation and comparison of structural variation. *F1000Research*. 2017;6:1795.
164. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, Shamardina O,

Stirrups K, Delon I, Dewhurst E, Dolling H, Erwood M, Grozeva D, Stefanucci L, Arno G, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome medicine*. 2018;10(1):95.

165. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, Montgomery SB, Wheeler M, Buchan JG, Lambert CC, Eng KS, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in medicine: official journal of the American College of Medical Genetics*. 2018;20(1):159–163.
166. Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, Nakamura M, Nagasaki M, Kinoshita K, Okamura Y, Miyatake S, Miyake N, Matsumoto N. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *Journal of human genetics*. 2019;64(5):359–368.